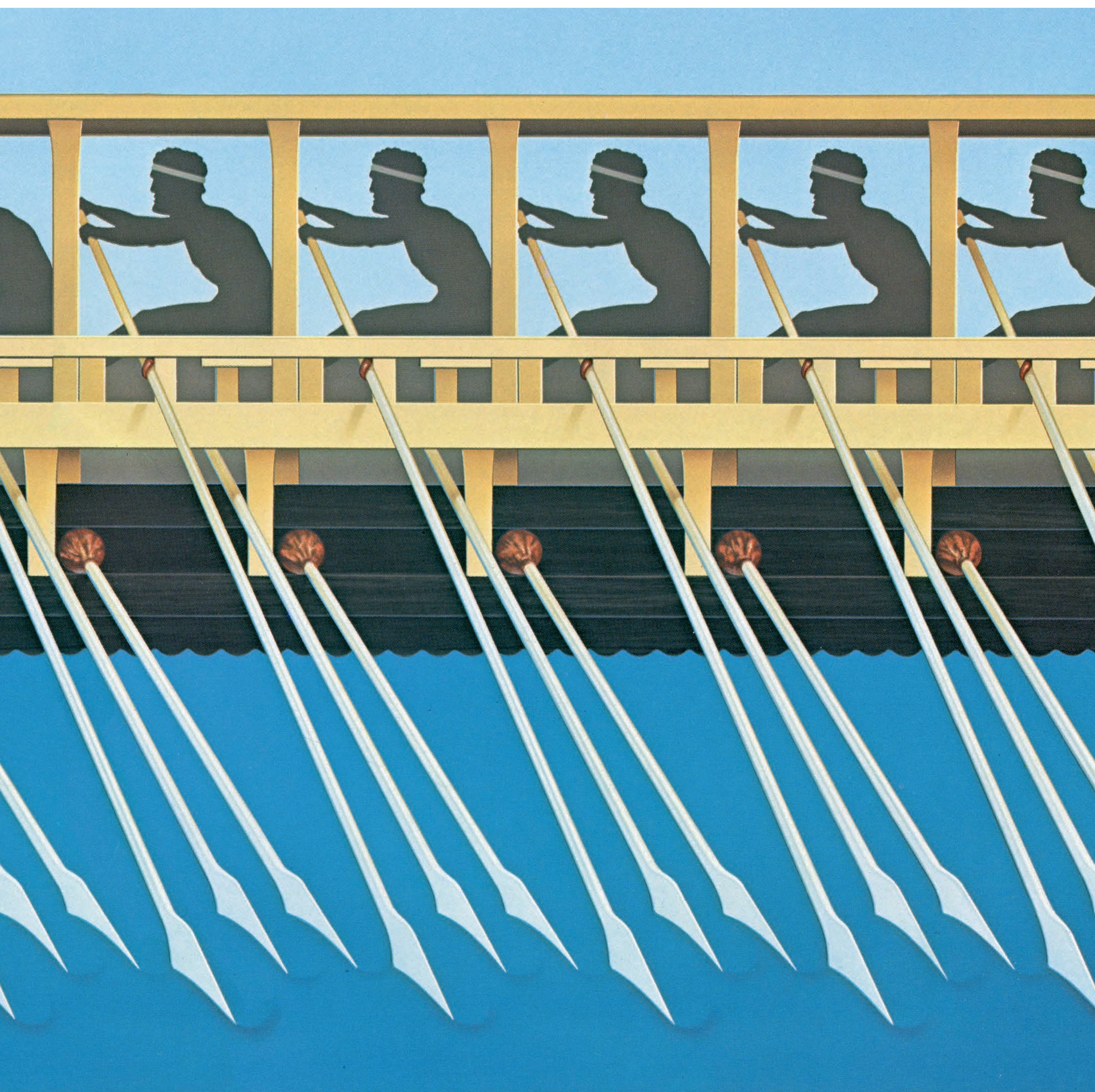
INVESTIGACION Y CIENCIA

Edición en español de SCIENTIFIC AMERICAN



NAVES DE GUERRA A REMO

Junio 1981

250 PTAS.

Los espacios en gris
corresponden a publicidad
en la edición impresa

Número 57
Junio 1981

INVESTIGACION Y
CIENCIA
Edición en español de SCIENTIFIC AMERICAN

Colaboradores de este número:

Asesoramiento y traducción:

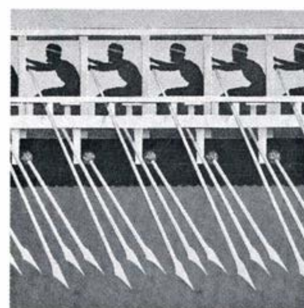
Antonio Travesí: *Liberaciones catastróficas de radiactividad;* Pedro Pascual: *Teoría unificada de las partículas elementales y las fuerzas;* Ramón Cerdà: *Reconocimiento del habla por medio de ordenadores;* Enrique Cerdá: *El origen de la información genética;* Manuel Puigcerver: *Las envolturas de las novas;* Joandomènec Ros: *Insectos filtradores;* Laureano Carbonell: *Naves de guerra a remo en la antigüedad;* Luis Bou: *Juegos matemáticos;* J. Vilardell: *Taller y laboratorio.*

Ciencia y sociedad:

Antonio Blanco, Julio Samsó, P. Cuñat, A. Aguilar y V. García

Libros:

Antonio Tordera, Josep Maria Tous y Luis Alonso



LA PORTADA

La ilustración de la portada representa una imagen ideal de una parte del costado de estribor de un trirreme, la nave de guerra empleada normalmente por las ciudades-estado griegas durante los siglos V y IV a. de C. (véase "Naves de guerra a remo en la antigüedad", de Vernard Foley y Werner Soedel). La nave ilustrada llevaba un total de 170 remeros distribuidos en tres órdenes, o niveles, distintos; cada uno de ellos manejaba un solo remo. Aquí se aprecian sólo los del orden situado más arriba; los remos se apoyaban en una postiza dispuesta por fuera del casco y a unos sesenta centímetros de distancia de los costados. Los remeros de los dos órdenes restantes estaban sentados en el interior del casco y un poco más bajos que aquéllos. Los remos de los que ocupaban el orden o nivel intermedio se apoyaban en la regala de la nave, es decir, en la tabla más alta del forro exterior, mientras los del orden más bajo pasaban por unos agujeros o portas practicados en el costado, recubiertos con un trozo de piel.

Suscripciones:
Prensa Científica, S. A.
Calabria, 235-239
Barcelona-29 (España)
Teléfono 322 05 51 ext. 41

Condiciones de suscripción:
España:
Un año (12 números): 2.750 pesetas
Extranjero:
Un año (12 números): 43 U.S.$
Ejemplar atrasado ordinario:
280 pesetas
Ejemplar atrasado extraordinario:
420 pesetas

Distribución para España
Distribuciones de Enlace, S. A.
Ausias March, 49, Barcelona-10

Distribución para los restantes países:
Editorial Labor, S. A.
Calabria, 235-239 - Barcelona-29

Publicidad:
Madrid:
Gustavo Martínez Ovin
Avda. de Moratalaz, 137, Madrid-30
Tel. 430 84 81
Cataluña:
Miguel Munill
Balmes, 191, 2.°, 2.ª, Barcelona-6
Tels. 218 44 45 y 218 40 86

Controlado por O.J.D.



PROCEDENCIA DE LAS ILUSTRACIONES

Pintura de la portada de George V. Kelvin

Página	Fuente	Página	Fuente
9-14	Alan D. Iselin	101	Douglas A. M. Craig, Universidad de Alberta
19-34	Gabor Kiss	102	J. Bruce Wallace, Universidad de Georgia
39	Stephen E. Levinson y Mark Y. Liberman	105	Lionel Casson, Universidad de Nueva York
40-47	Jerome Kuhl	106-117	George V. Kelvin
48	Stephen E. Levinson y Mark Y. Liberman	118	Lionel Casson
49-51	Jerome Kuhl	123-125	M. Alonso, S. Grisolía, E. Knecht y J. Hernández-Yago
63-81	Allen Beechel	127-128	S. Grisolía, E. Knecht y J. Hernández-Yago
83	Walken Graphics	129	M. Alonso, S. Grisolía, E. Knecht y J. Hernández-Yago
84	Observatorio Steward	130	S. Grisolía, E. Knecht y J. Hernández-Yago
85	Observatorio Yerkes (*arriba y en el centro*), Observatorio Lick (*abajo*)	131	M. Alonso, S. Grisolía, E. Knecht y J. Hernández-Yago
86-90	Walken Graphics	132	S. Grisolía, E. Knecht y J. Hernández-Yago
91	Observatorio Lick	133-137	M. Alonso, S. Grisolía, E. Knecht y J. Hernández-Yago
95	Douglas A. M. Craig, Universidad de Alberta (*arriba y abajo, a la derecha*); J. Bruce Wallace, Universidad de Georgia (*abajo, a la izquierda*)	138-140	Ilil Arbel
96-97	Tom Prentiss	143	R. G. Olsson E. T. Turkdogan
98	J. Bruce Wallace, Universidad de Georgia	144-148	Michael Goodman
99-100	Tom Prentiss		

ISSN 0210-136X
Dep. legal: B. 38.999-76
Fotocomposición Tecfa
Pedro IV, 160 - Barcelona-5
Fotocromos reproducidos por GINSA, S.A.
Imprime GRAFESA
Gráfica Elzeviriana, S. A.
Nápoles, 249 - Tel. 207 40 11
Barcelona-13
Printed in Spain - Impreso en España

Los autores

STEVEN A. FETTER y KOSTA TSIPIS ("Liberaciones catastróficas de radiactividad") se hallan adscritos al programa de ciencia y tecnología para la seguridad internacional del departamento de física del Instituto de Tecnología de Massachusetts. Fetter está a punto de terminar sus estudios universitarios. Tsipis ostenta el cargo de director adjunto del programa de ciencia y tecnología. Griego de origen, se trasladó en 1974 a los Estados Unidos para estudiar ingeniería eléctrica y física. Licenciado por la Universidad de Rutgers, se doctoró por la de Columbia. Está en el departamento de física del MIT desde 1966.

HOWARD GEORGI ("Teoría unificada de las partículas elementales y de las fuerzas") enseña física en la Universidad de Harvard. Graduado en el Harvard College, recibió el doctorado por la Universidad de Yale en 1971. Volvió luego a la Universidad de Harvard, donde disfrutó de varias becas postdoctorales antes de entrar en el claustro de la facultad, en 1976.

STEPHEN E. LEVINSON y MARK Y. LIBERMAN ("Reconocimiento del habla por medio de ordenadores") pertenecen a la plantilla técnica de los Laboratorios Bell. Levinson se licenció en ingeniería técnica por el Harvard College, doctorándose en la especialidad de eléctrica por la Universidad de Rhode Island. Antes de ingresar, en 1976, en los Laboratorios Bell, enseñó cibernética durante algunos años en la Universidad de Yale. Liberman es doctor en lingüística por el Instituto de Tecnología de Massachusetts.

MANFRED EIGEN, WILLIAM GARDINER, P. SCHUSTER y RUTHILD WINKLER-OSWATITSCH ("El origen de la información genética") prepararon su artículo en el Instituto Max Planck de Química Biofísica en Göttingen, donde Eigen dirige el departamento de cinética bioquímica. Además de su trabajo teórico sobre evolución molecular, Eigen continúa su investigación experimental sobre diversos problemas de la cinética de las reacciones, empleando los métodos de relajación química que le valieron el Nobel en 1967. Gardiner, profesor de química en la Universidad de Texas en Austin, llegó al campo de la evolución molecular tras estudios experimentales y simulación en ordenador de cinética de gases y explosiones. Schuster, doctorado por la Universidad de Viena en 1967, es profesor y director del Instituto de Química Teórica y Radioquímica de esa universidad. Combina su interés por la dinámica de las reacciones con estudios de los puentes de hidrógeno. Winkler-Oswatitsch adquirió su doctorado en la Universidad Técnica de Viena en 1969. Sus intereses como investigadora van desde la cinética de las reacciones de iònes con antibióticos hasta problemas evolutivos y el análisis filogenético del ARN.

ROBERT E. WILLIAMS ("Las envolturas de las novas") es profesor de astronomía en la Universidad de Arizona y astrónomo del Observatorio Steward de dicho centro superior. Comenzó sus estudios en la Universidad de California en Berkeley, terminándolos en la de Wisconsin en Madison, donde obtuvo el doctorado en 1965. Desde entonces ha sido miembro del cuerpo docente de la Universidad de Arizona. Los temas de investigación que interesan a Williams, aparte de las envolturas de las novas, son el análisis de la radiación procedente de los quasars y la estructura de los discos de acumulación en sistemas binarios de estrellas.

RICHARD W. MERRITT y J. BRUCE WALLACE ("Insectos filtradores") son entomólogos que comparten un interés común por la ecología de los insectos que viven en ríos y corrientes de agua. Merritt es profesor adjunto de entomología en la Universidad estatal de Michigan. Durante su formación pasó por la Universidad estatal de California en San José, estatal de Washington y Universidad de California en Berkeley, donde se recibió de doctor en 1974. De su tesis doctoral nos comenta: "trataba del estudio ecológico del complejo de insectos asociado con las deyecciones del ganado. Aunque mi atención principal se centra ahora en la ecología fluvial del estado de Michigan, hay muchas semejanzas entre los dos hábitats con respecto a la fauna de insectos que en ellos vive". Wallace da clases de entomología en el Instituto de Ecología de la Universidad de Georgia. Se graduó en la Universidad Clemson y alcanzó el grado de doctor por la estatal de Virginia en 1967. Está investigando la respuesta de los ecosistemas acuáticos a las alteraciones de la cuenca hidrográfica.

VERNARD FOLEY y WERNER SOEDEL ("Naves de guerra a remo en la antigüedad") han colaborado en una gran variedad de materias relacionadas con la historia de la ciencia y de la técnica, entre ellas, la investigación sobre el funcionamiento de las catapultas usadas en la antigüedad. Sobre este tema en particular publicaron, en el número de mayo de 1979 de nuestra revista, el artículo titulado "Catapultas antiguas". Ambos pertenecen a la Universidad de Purdue. Vernard Foley, profesor adjunto de historia, procede del MacPherson College y de las Universidades de Kansas y de California en Berkeley. Actualmente está investigando la participación de Leonardo da Vinci en el desarrollo de la fresadora. Werner Soedel, por su parte, es profesor de ingenirería mecánica. Nacido en Praga, se educó en Alemania Occidental, donde obtuvo su título de ingeniería. Doctorado en Purdue, está interesado en la historia de la técnica e investiga los fenómenos relacionados con la vibración.

SANTIAGO GRISOLIA, ERWIN KNECHT y JOSE HERNANDEZ-YAGO ("Proteolisis intracelular") comparten en los últimos años tareas de investigación en el campo de las vías, mecanismos y regulación de la degradación intracelular de proteínas en el Instituto de Investigaciones Citológicas de la Caja de Ahorros de Valencia. El doctor Grisolía es profesor distinguido Sam E. Roberts de bioquímica y biología molecular en la Universidad de Kansas. Gran parte de su carrera científica ha estado dedicada a la enzimología. Desde 1945 ha desarrollado su labor científica en los Estados Unidos. Tiene numerosas publicaciones, especialmente sobre el metabolismo del fosfoglicerato, metabolismo de carbamil- y acetil- aminoácidos e inactivación de enzimas inducida por sustratos. Desde 1977 ostenta el cargo de director del Instituto de Investigaciones Citológicas. Hernández-Yago (doctor ingeniero agrónomo) es jefe del departamento de microscopía electrónica del Instituto de Investigaciones Citológicas de Valencia desde 1972 y, actualmente, subdirector del centro. Director de los cursos para postgraduados que anualmente se celebran en el Instituto sobre técnicas de microscopía electrónica en biología, tiene publicadas monografías y numerosos trabajos en el campo de la biología celular. Knecht (doctor en ciencias biológicas) es, desde 1974, adjunto de investigación del Instituto de Investigaciones Citológicas.

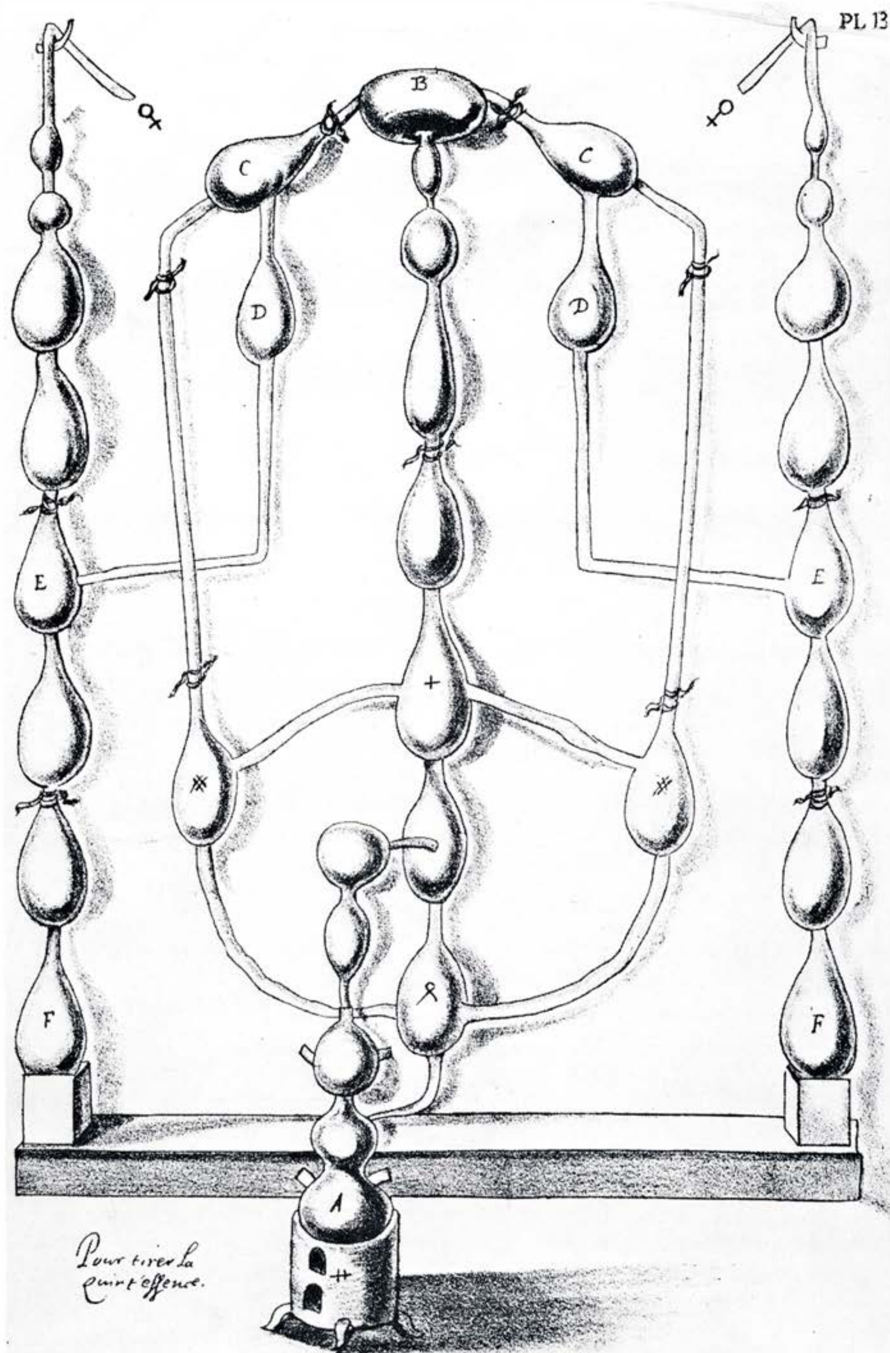
Hace…

José M.ª López Piñero

… cuatrocientos años

Adquirió importancia la actividad desarrollada en el laboratorio de destilación anejo a la "botica" del Monasterio del Escorial. En su *Historia de la Orden de San Jerónimo* (1605), José de Sigüenza expone que fue construido por iniciativa personal de Felipe II y habla con admiración de los aparatos instalados en sus siete u ocho habitaciones, "con que se hacen mil pruebas de la naturaleza y que con la fuerza del arte y del fuego y otros medios e instrumentos descubren sus entrañas y secretos". Su testimonio es el de un profano que ve "pruebas de cosas maravillosas", pero resulta claro que allí se obtenían "quintaesencias y aceites" de muy diferentes vegetales y minerales, así como preparados alquímicos, entre ellos el llamado "oro potable". Parecida es la actitud de Jerónimo de Sepúlveda: "¿A quién no admiran aquellas máquinas tan grandes de sacar aguas por vidrio? ¡Qué de cosas preciosas y de gran valor hay en esta oficina!". Muy distinta es la información que proporciona Jehan Lhermite, gentilhombre de cámara de Felipe II, que residió en España entre 1586 y 1602. En su obra *Le Passetemps* describe con cierto detalle la "mayson pour distiller des eaux", como un edificio cercano pero independiente de la "botica" propiamente dicha. Reproduce una lista pormenorizada de los productos que allí se obtenían, que encabeza el famoso "oro potable" alquímico, pero que en su mayor parte está integrada por "aguas destiladas de toda clase de hierbas, metales y especies" y "quintaesencias". Las actividades fundamentales del laboratorio eran, en efecto, la preparación de medicamentos y la obtención de perfumes. Detalla también el funcionamiento de sus tres principales aparatos, adjuntando incluso dibujos de los mismos. El primero es el utilizado para obtener las quintaesencias. Consta nada menos que de veintiséis "vasos de vidrio, unidos entre sí con largos tubos también de vidrio". El calor se aplica únicamente en el horno sobre el que descansa el primero de ellos, "donde se coloca la materia de la que se pretende extraer la quintaesencia". El segundo aparato es la llamada "torre filosofal" y es "el principal instrumento para destilar aguas de toda clase, en abundancia". Tiene una altura de unos veinte pies y un diámetro tal "que tres hombres apenas la pueden abrazar". Además del horno y la base de ladrillos, "está hecha de latón, en forma de torre, y destila por el calor de vapor; contiene un gran número de vasos o alambiques de vidrio, y en veinticuatro horas extrae más de 200 libras de aguas destiladadas de la clase de hierbas que en ella se colocan". El tercer aparato es el "destilatorio" de vapor ideado por Diego de Santiago al que a continuación nos referiremos.

¿Quiénes trabajaron en este laboratorio? Por supuesto, una serie de boticarios, entre ellos fray Jerónimo de Albendea, maestro oficial de la "botica", del que habla Jerónimo de Sepúlveda y Juan del Castillo, boticario de origen francés residente en Cádiz, que se había formado allí. En su *Pharmacopoea Universa Medicamenta* (1622), Castillo

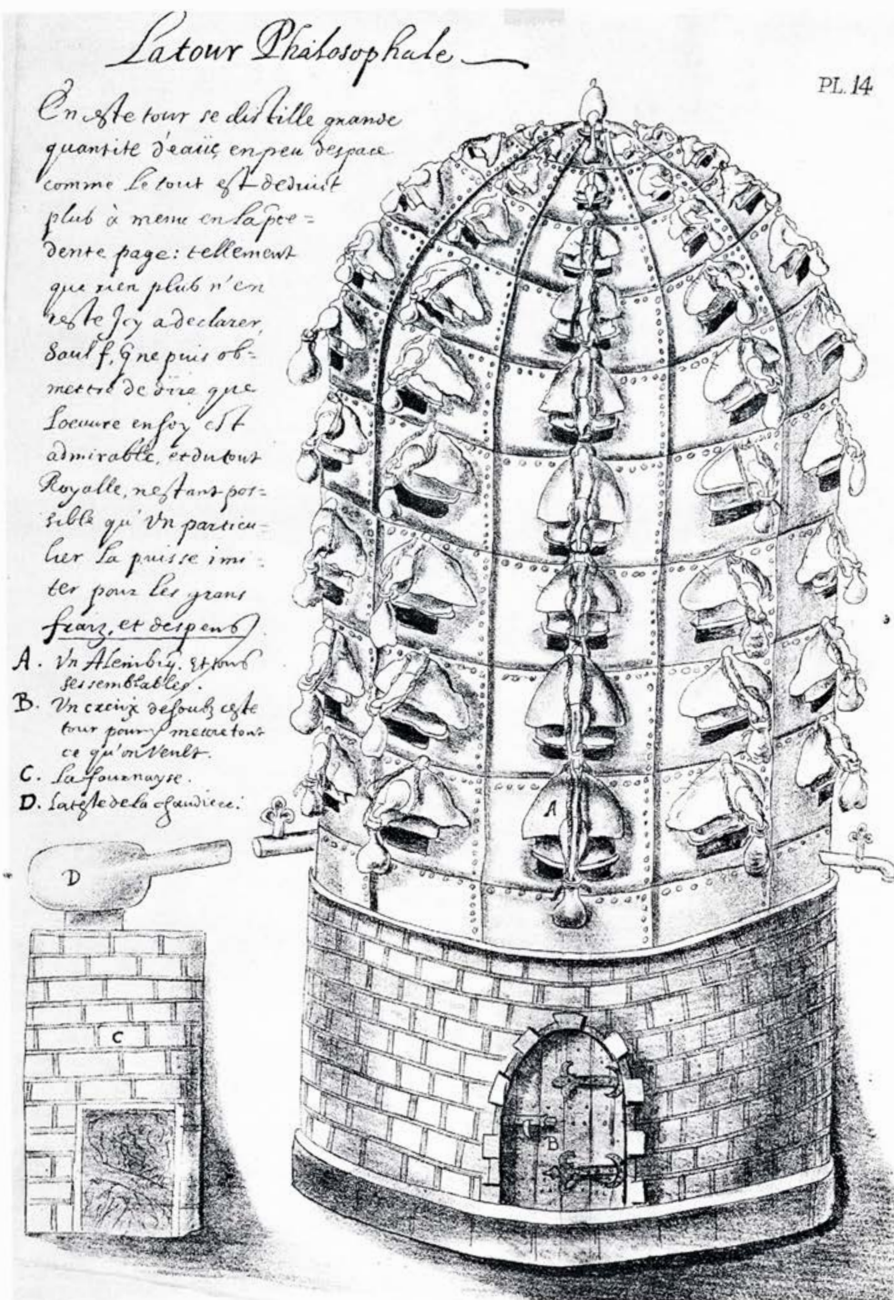


Las ilustraciones de esta sección son dibujos de **Le Passetemps,** *de Jehan Lhermite, que representan los tres principales aparatos instalados en el laboratorio de destilación del Monasterio del Escorial a finales del siglo* XVI. *Esta lámina, la 13, representa un aparato para obtener quintaesencias*

trata con cierta amplitud de la destilación y afirma que sus procedimientos, "si no se ven (...) con mucha dificultad las harán (...) y viéndolo lo aprenderán más presto que por dicho escrito, y para quintas esencias al Escurial en la botica de San Jerónimo".

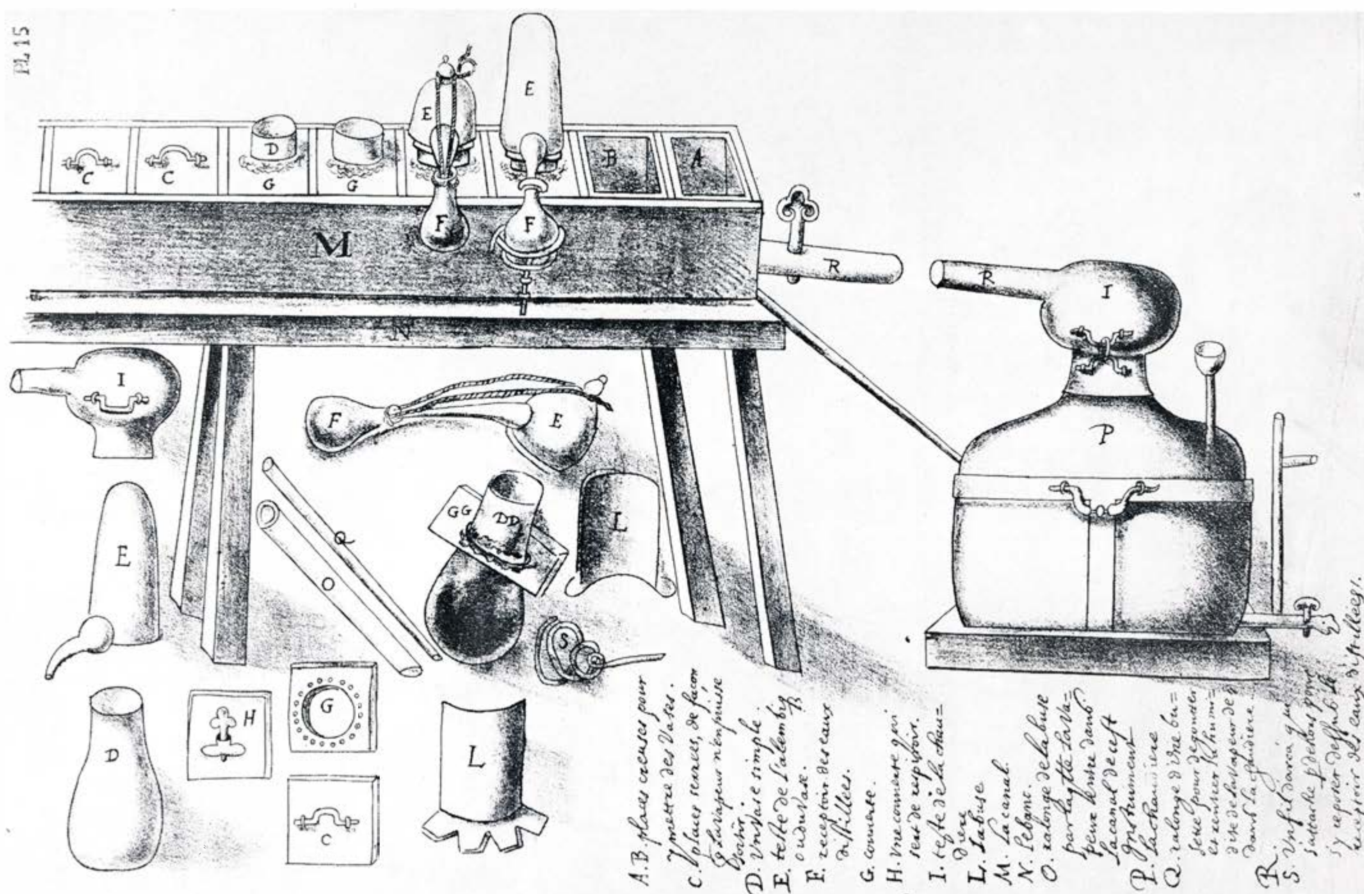
También trabajaron en el laboratorio diferentes alquimistas. Existen pruebas documentales de que en 1557, 1559 y 1567 varios "maestros" habían trabajado al servicio de Felipe II. En la parte final del siglo, una vez montado el laboratorio, sabemos igualmente que trabajaba en El Escorial un tal Richard Stanihurst, que dedicó al monarca en 1593 una obra titulada *El toque de Alchimia.* Se trata de una exposición, que quedó manuscrita, destinada a "declarar los verdaderos y falsos efectos del arte (alquímico) y cómo se conoscerán las falsas prácticas de los engañadores y haraneros vagamundos".

En conexión con este ambiente estaba asimismo el boloñés Leonardo Fioravanti, sobradamente conocido como el principal paracelsista italiano. La relación de Fioravanti con España procedía de sus años de Nápoles, en los que se convirtió en el médico preferido de los gobernantes españoles de aquel territorio. Se reunían ya entonces a practicar en su casa "alchimisti di diverse nationi". En 1551 el virrey de Nápoles Pedro de Toledo le nombró médico de cámara de su hijo García de Toledo y con éste salió para Africa en la flota del emperador Carlos V. Años más tarde dedicó a Felipe II su obra titulada *Della Fisica* (1592), cuyo libro IV está consagrado a la alquimia. Este libro nos permite reconstruir interesantes detalles de su estancia en España durante los años 1576 y 1577. Como buen paracelsista, Fioravanti ocupó una posición intermedia entre la ciencia académica y la alquimia extraacadémica. Por ello, durante su estancia en nuestro país se movió también en un nivel intermedio entre ambas. Trató con numerosos científicos y médicos y hace, por ejemplo, grandes elogios de Monardes. Pero, según propia declaración, tanto en Madrid como en Barcelona y Navarra le consideraron unas veces un "gran médico" y otras un "alquimista" y un "nigromante". De hecho mantuvo también estrecha relación con varios alquimistas españoles, intercambiando con ellos toda clase de noticias. Reproduce por ello al final de su *Física* el texto de las *Coplas sobre la piedra philosophal* del valenciano Luis de Centelles, que le proporcionó un alquimista madrileño. Cabe pensar que Fioravanti contribuyó a la difusión entre los alquimistas españoles de las obras de Paracelso que acabaron pasando a primer plano en la literatura por ellos manejada.



La llamada "torre filosofal" (lámina 14)

Los principales encargados del laboratorio del Escorial eran, sin embargo, los "destiladores de Su Majestad". Se trata de uno de los numerosos puestos de carácter científico o técnico que figuraban en la casa real en tiempos de Felipe II. El nombramiento más antiguo del que tenemos noticia (1572) corresponde a Francisco Holbecq, hijo de un jardinero flamenco que trabajaba al servicio del monarca. Como puso de relieve G. de Amezúa, simultaneó su oficio de "destilador de aguas y aceites" con la supervisión de los jardines reales. Poco más tarde, los "destiladores de Su Majestad" fueron un grupo integrado por extranjeros como el propio Holbecq o el italiano Juan Vicencio Forle y por españoles como Juan del Valle y Diego de Santiago. Este último no solamente fue la personalidad científica más destacada que trabajó en el laboratorio, sino la principal figura española de la etapa previa a la constitución de la química. Su obra *Arte separatoria y modo de apartar todos los Licores, que se sacan por vía de Destilación* (1598) no es una monografía más sobre destilación, sino un escrito que desde muchos puntos de vista tiene interés para la historia de la química europea. Ignorada por los historiadores extranjeros, incluso en los estudios consagrados a la historia de la destilación, no ha recibido la debida atención por parte de los historiadores españoles.



El "destilatorio" de vapor ideado por Diego de Santiago (lámina 15)

Diego de Santiago, nacido a mediados del siglo XVI en una pequeña localidad de Extremadura, pasó toda su vida dedicado al trabajo de laboratorio en su pueblo natal, en Zamora, en El Escorial al servicio del rey y en Sevilla, ciudad en la que residía cuando publicó su libro. Incluye éste un detallado estudio de los instrumentos, técnicas y materiales empleados en la destilación, un resumen de sus fundamentos teóricos y una amplia exposición de sus aplicaciones a la preparación de medicamentos y también a cuestiones relacionadas con las conservas, los vinos, el análisis de las aguas, los venenos, etc.

La obra de Santiago puede servir de ejemplo destacado del paso a primer plano de la experimentación como criterio científico y, en consecuencia, de rechazo especialmente duro de la autoridad de los clásicos desde una clara idea del progreso científico. Su libro corresponde a los resultados de toda una vida de trabajo, "en especial de veinte años a esta parte, comunicándolos con los Destiladores de Su Majestad, confiriendo con médicos y siempre haciendo experiencias, en las cuales y en varios instrumentos que he inventado se ha gastado cuanto mi trabajo me ha podido dar". Resulta lógico que el texto carezca casi totalmente de citas, ya que "cuando la cosa se ve, no tenemos necesidad de autoridades ni alegaciones". Por ello, adquiere mayor relieve la única referencia que aparece en todo el libro: "Los que siguen la doctrina de los antiguos, cuando se ofrece alguna ocasión de tratar los efectos de las medicinas espirituosas, que los dichos antiguos no conocieron (esta fue la causa de no tratar dellas) y por haberlas ellos ignorado, no quieren creer lo que dellas dicen los modernos; los cuales con muy justas causas han venido a tener el dicho conocimiento, el cual se alcanza por medio de nuestra arte separatoria, la cual entendieron muy bien Arnaldo de Vilanova y Raymundo Lulio y Theophrasto Paracelso y Vbequero y Joannes de Rupecissa, y otros muchos que han seguido el arte separatoria, por cuyos medios han venido a sacar a luz lo oculto de la naturaleza; con lo cual se hacen los efectos que ignoran los que siguen la medicina corporal". La tradición en la que se inscribe Santiago es bien clara: los tres grandes nombres de la alquimia bajomedieval, Wecker como típico representante de la fase empírica correspondiente a la literatura "de secretis" y Paracelso. La influencia de este último, está asimilada por un científico de talante crítico, de modernidad a menudo sorprendente, aunque sean también evidentes algunos rasgos que lo relacionan con la cultura extraacadémica de los alquimistas. Santiago habla no obstante ya de "arte separatoria" y la influencia que sobre él ejerce Paracelso no queda reducida a un mero complemento de las ideas tradicionales, sino que sirve para contraponer orgullosamente la medicina "de los modernos" con la "medicina antigua". En sus ataques a ésta, insiste en la cuestión central del método: "La medicina antigua debe haber sido escripta, discurriendo con el entendimiento, sin venir a la demostración y experiencia".

Portela, al que se debe el primer análisis de la obra de Santiago, lo considera como el texto químico de mayor importancia de la España del siglo XVI y el que mejor representa "la confluencia del paracelsismo y la alquimia, dentro de un marco de estricta modernidad". Ha estudiado con detenimiento el "instrumento separatorio" inventado por el extremeño, "el mejor y más fácil que hasta hoy se ha hallado". Es un "destilatorio de vapor" que consta de una serie de vasos de vidrio acoplados en un cuadro metálico, que se suspenden en un canal de barro o de cobre que actúa como transportador del vapor generado en una caldera.

Liberaciones catastróficas de radiactividad

El accidente más grave concebible en un reactor nuclear es bastante menos destructivo que la detonación de un arma nuclear, incluso considerando únicamente los daños que ésta produce por radiación

Steven A. Fetter y Kosta Tsipis

De múltiples formas una fracción muy importante de la población humana puede verse expuesta a cantidades peligrosas de radiactividad: a consecuencia, inevitable, de una guerra nuclear, por limitada que se la suponga; tras un accidente en un reactor nuclear que causase la explosión de la vasija de contención, lo que permitiría que el material del núcleo del reactor escapase a la atmósfera; por descarga inadvertida de agua o de gases de un reactor que porten núclidos radiactivos, creando el peligro de una exposición a la radiación de una magnitud menor; y, por último, a raíz de un accidente en las fases de fabricación, transporte, reprocesado o almacenamiento de materiales radiactivos para reactores o armas nucleares.

Difiere notablemente, de un suceso a otro, la cantidad de radiactividad que se liberaría; por tanto, cada posibilidad de éstas debe considerarse por separado. Describiremos la radiactividad que probablemente se liberaría en cada uno de los tres casos siguientes: explosión o detonación de un arma nuclear en el suelo, fusión del núcleo de un reactor nuclear y rotura, por explosión, de su vasija de contención con el resultado de un escape de radiactividad y, en tercer lugar, explosión de una cabeza termonuclear sobre un reactor nuclear.

No incluiremos en estas comparaciones la onda de choque ni el calor que constituye el efecto explosivo inmediato de un arma de guerra termonuclear. Tan sólo examinaremos y compararemos los efectos retardados que produce la liberación de radiactividad. Destaca claramente el hecho de que la detonación de un arma nuclear es mucho más temible que cualquier tipo de accidente que pueda registrarse en un reactor nuclear. Sin embargo, la explosión de un arma nuclear sobre un reactor es mucho más dañina que la detonación de este arma sobre el suelo en cualquier otro lugar. El ataque nuclear convierte un reactor en un arma radiológica devastadora.

Las armas termonucleares suelen constar de tres partes. El primer componente viene a ser un detonador, cuyo elemento más importante son unos kilos de plutonio. La fisión de los núcleos de plutonio origina el calor necesario para que se produzca una explosión termonuclear.

La segunda parte está constituida por el explosivo termonuclear, una mezcla de deuterio y tritio, isótopos pesados del hidrógeno. La fusión termonuclear de un núcleo de deuterio (que tiene un neutrón y un protón) con un núcleo de tritio (que tiene un protón y dos neutrones) produce un núcleo de helio (con dos protones y dos neutrones). El neutrón sobrante es emitido con alta velocidad y se libera, en forma de calor, una gran cantidad de energía. Los productos de esta reacción no portan radiactividad de larga duración.

La tercera parte del arma termonuclear es una capa de uranio que rodea la masa de deuterio y tritio. Los núcleos de los átomos de uranio se fisionan cuando son bombardeados por los neutrones emitidos por la fusión termonuclear. Los fragmentos de los núcleos fisionados constituyen una fuente abundante de radiactividad. En un arma nuclear como ésta, casi la mitad de la energía liberada proviene de la fusión termonuclear y la otra mitad de la fisión del uranio.

El calor generado por la detonación de un arma termonuclear vaporiza el ingenio casi instantáneamente, cesando las reacciones nucleares. La mayoría de los núcleos creados por la fisión del uranio quedan en un estado de energía anormalmente elevado. Su transición a un estado de energía más bajo se realiza por emisión de radiación en la zona de alta energía del espectro electromagnético, que corresponde a los rayos X y a la radiación gamma. Esta radiación calienta el aire de los alrededores formando una onda de choque que va elevando la temperatura de capas adicionales de aire. Resulta así una bola de fuego luminosa. En un ingenio nuclear con una potencia explosiva de un megatón (la energía equivalente a un millón de toneladas de explosivos químicos), la bola de fuego se eleva a una velocidad de 120 metros por segundo hasta una altitud de unos 18.000 metros. Un megatón es la potencia típica de una cabeza nuclear de un misil balístico intercontinental en el arsenal de la Unión Soviética.

La corriente ascendente generada por la bola de fuego arrastra grandes cantidades de tierra y residuos. Una explosión de un megatón en la superficie del terreno puede excavar un cráter de más de 365 metros de diámetro y 120 metros de profundidad. Cuando la bola de fuego se enfría, los núcleos radiactivos creados por la explosión se condensan dentro de las partículas de tierra y cenizas que en el transcurso del tiempo vuelven a la tierra en forma de lluvia radiactiva.

Esta lluvia es radiactiva, en parte, porque algunos de los núcleos creados por la explosión son inestables y tienen en general un exceso de neutrones. El remedio para esta inestabilidad es que un neutrón se transforme en un protón por el proceso denominado desintegración beta. En el curso de este proceso, un núcleo expulsa un electrón, al que en este contexto se denomina rayo be-

ta. Tal transformación puede dejar el núcleo, a su vez, en un estado excitado, del que pasa a su nivel fundamental o de mínima energía emitiendo radiación electromagnética, principalmente rayos gamma. Las partículas de la lluvia radiactiva continúan emitiendo rayos beta y gamma durante muchas décadas después de la explosión. Estas emisiones son completamente aleatorias. Dada una cantidad determinada de núclidos radiactivos, únicamente se puede predecir el número promedio de desintegraciones en un intervalo determinado de tiempo. Con el paso del tiempo, el número de núcleos en estado inestable o excitado decrece de modo que la intensidad de la radiactividad disminuye.

Se emplean varias unidades de medida para describir la cantidad de radiactividad o la cantidad de energía que la radiactividad puede depositar en los tejidos vivos. La unidad estándar de la radiactividad como tal es el curie, que se define como $3,7 \times 10^{10}$ emisiones o desintegraciones por segundo. Esta definición no hace referencia, empero, al tipo de radiación ni a su energía. La unidad que mide la energía depositada por la radiactividad en un medio material es el rad, que se define como la absorción de 100 erg por gramo de materia, el tejido vivo por ejemplo. Otra unidad es el roentgen, que se refiere



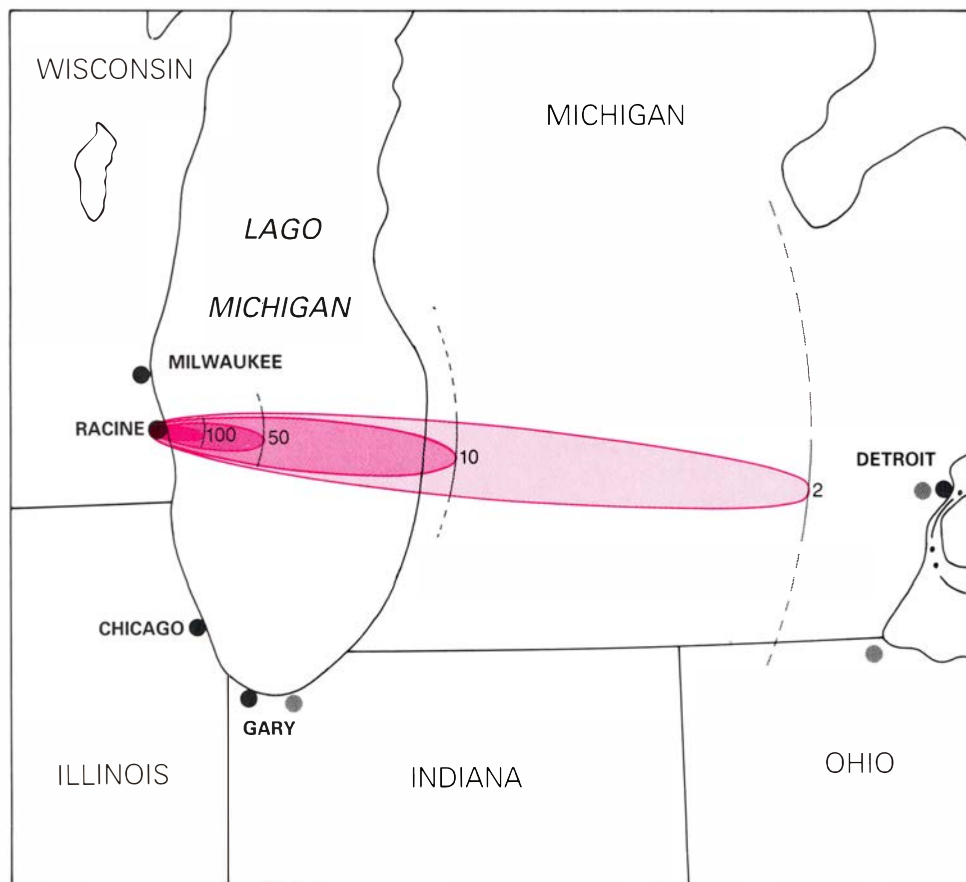
ATAQUE A UN SOLO REACTOR con una única bomba termonuclear capaz de devastar una parte substancial de Europa. Aquí se ha realizado un hipotético ataque al reactor nuclear de un gigawatt localizado en Neckarwestheim, en Alemania Occidental. Se supone que la cabeza nuclear tiene un megatón de potencia. El viento predominante es del sureste, con una velocidad de 25 kilómetros por hora. Un mes después del ataque, la zona donde la dosis es de 10 rem por hora (*en color claro*) puede extenderse hasta muy al interior del Reino Unido. Un año después del ataque, la zona de 10 rem por año todavía abarca (*color oscuro*) la mayor parte de la capacidad industrial de Alemania Occidental. Los puntos grises señalan la ubicación de reactores nucleares comerciales.

exclusivamente a los rayos X y gamma. La exposición a un roentgen de rayos gamma equivale a la absorción de 94 erg por gramo de tejido. De ello se deduce que el rad y el roentgen vienen a ser prácticamente equivalentes.

Dado que ninguna de estas unidades describe la cantidad del daño biológico producido por la radiación, se precisa una tercera unidad todavía. Se trata del rem, una abreviación de las palabras inglesas "roentgen equivalent man" (roentgen equivalente en el hombre). Una dosis de radiación medida en rem tiene en cuenta el hecho de que los distintos tipos de radiaciones pueden desarrollar efectos completamente diferentes en un organismo vivo, aun cuando se deposite la misma cantidad de energía y el daño se produzca por el mismo mecanismo general, o sea, la ionización de los átomos en las moléculas intracelulares. Las diferencias en el daño producido reflejan fundamentalmente ciertas características de la radiación: su poder de penetración o distancia hasta donde penetra en un tejido dado. Una dosis en rem es igual a una dosis en rad multiplicada por un factor llamado eficiencia biológica relativa (EBR), específico de cada tipo particular de radiación. Para las radiaciones beta y gamma, el EBR es aproximadamente la unidad. En adelante, pues, una dosis en rad se considerará igual a una dosis en rem. Puede conseguirse cierto sentido del tamaño de las dosis comparándolas entre sí en el siguiente par de ejemplos: una radiografía de los pulmones significa una dosis de aproximadamente 0,01 rem absorbida en una fracción de segundo; la radiación natural de fondo a nivel del mar asciende a aproximadamente 0,075 rem por año.

Los efectos biológicos de la radiación varían considerablemente de una persona a otra. Dependen de condiciones tales como la edad y la salud del sujeto. Ello impide definir niveles precisos de radiación en los que pueda esperarse encontrar los síntomas del síndrome radiactivo: caída del cabello, vómitos, diarrea, hemorragias internas y lesiones en boca y garganta. Sin embargo, se ha determinado que si el cuerpo humano queda expuesto a más de 500 o 600 rem a lo largo de un intervalo de tiempo de uno o dos días, las posibilidades de supervivencia son prácticamente nulas. Si la dosis está entre 200 y 450 rem la supervivencia, aunque posible, no puede asegurarse ni siquiera con tratamiento médico. Con todo ello presente, parece razonable suponer que una dosis de 400 rem en un día acarrea un porcentaje de mortandad del 50 por ciento o superior. La exposición de una población a 100 rem en el mismo período produciría enfermedades y algunas muertes. Sin embargo, con este nivel de dosis podría esperarse que la mayoría de las personas se recuperasen incluso sin atención médica.

Para calcular la extensión de territorio hecha inhabitable por una liberación dada de radiactividad, tomaremos como dosis máxima aceptable la de 2 rem por año. Esta dosis es más de 10 veces superior a la dosis máxima recomendada por la Oficina de Protección del Medio Ambiente de los Estados Unidos, y más de 20 veces la dosis de la radiación natural de fondo. Pero tam-





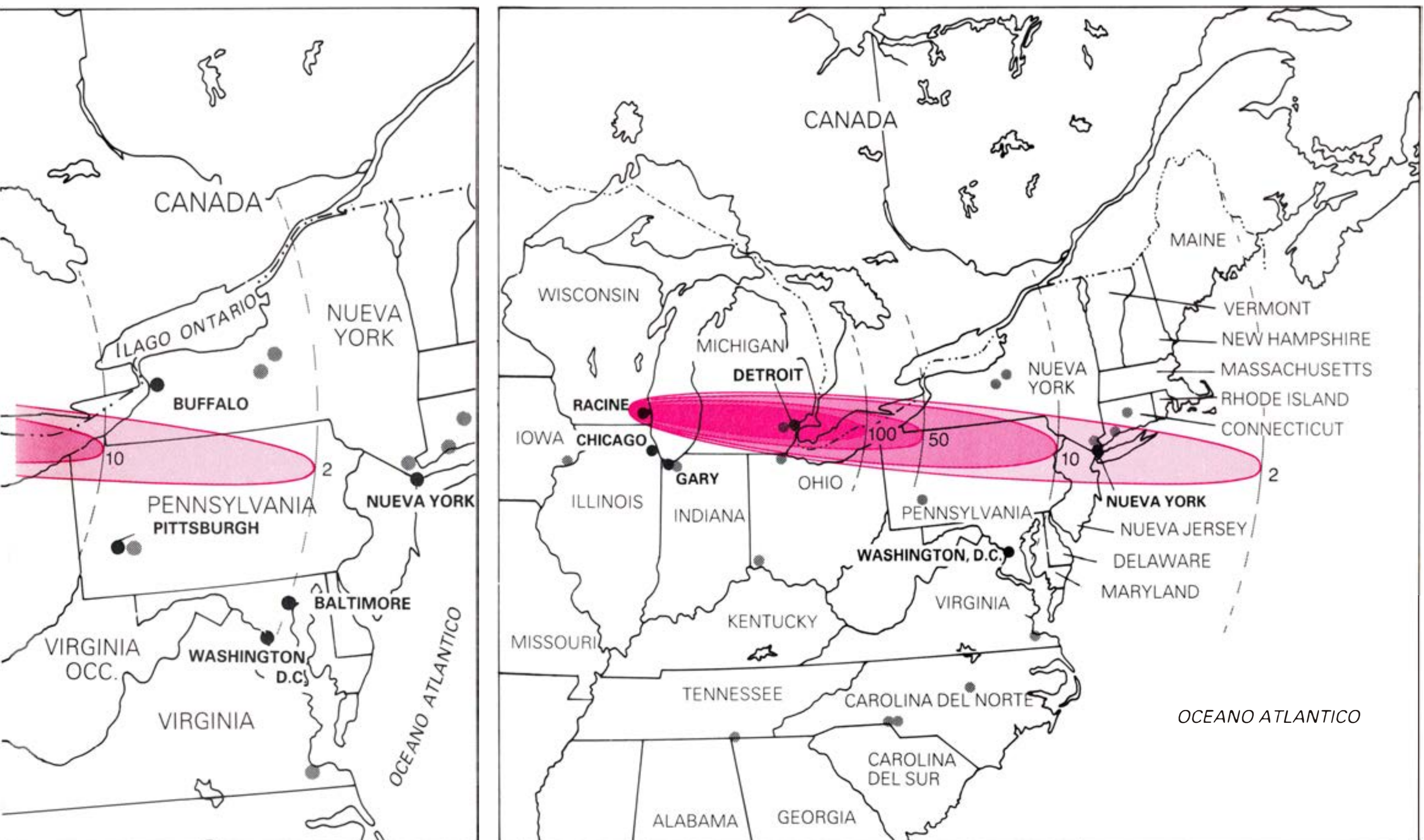
SE COMPARAN LOS CONTORNOS de las dosis producidas en tres hipotéticas liberaciones de radiactividad. El mapa de la izquierda recoge la forma que presenta la zona contaminada una semana después del accidente en el que el núcleo de un reactor nuclear de un gigawatt libera un tercio de su radiactividad. La cantidad de radiactividad liberada es cien millones de veces mayor que la que se liberó durante el accidente de la Three Mile Island cerca de Harrisburg, en 1979. El mapa central muestra la forma presentada por el área contaminada una semana después de que un arma nuclear de un megatón de potencia explote en la superficie del terreno en Racine, aunque sin dañar el reactor, y liberando inicialmente una cantidad mucho mayor de radiactividad. El mapa de la derecha señala cómo se comportaría el área de contaminación una semana después de que un arma nuclear de un megatón

bién es inferior a los 5 rem por año, que actualmente se considera como el límite superior para los trabajadores expuestos a la radiación durante años. Una dosis estándar de 2 rem por año podría adoptarse muy bien como límite en el período inmediato tras un accidente nuclear en tiempos de paz. Sin embargo, en el caso de una guerra nuclear, es muy poco probable que el público pueda ser evacuado de todas las zonas donde el nivel de radiación sea de 2 rem por año. Sin duda, la gente impulsada por el hambre u otras causas podría desear (o verse obligada) a asentarse en lugares que absorbieran más de 50 rem por año, una dosis que provoca la enfermedad de la radiación a más de la mitad de la población expuesta. Una dosis de 50 rem por año produce también víctimas ocasionales y tumores cancerosos en los individuos años después de la exposición.

Volvamos a las consecuencias de la liberación de radiactividad por una bomba termonuclear de un megatón detonada en la superficie del terreno. La mayor parte de la lluvia radiactiva resultante vuelve a la tierra en la dirección del viento predominante desde el momento de la explosión, y el 70 por ciento de la lluvia radiactiva lo constituyen partículas relativamente grandes que regresan a la tierra en el intervalo de un día. La intensidad de la radiación decrece con la distancia a partir del lugar de la explosión. Por un lado, la nube radiactiva de cenizas pierde partículas de polvo cuando el viento la arrastra y, por otro, la radiactividad disminuye a medida que se van desintegrando los núcleos radiactivos.

Con viento constante, las líneas de las dosis de radiactividad (en rem) acumuladas dibujan un conjunto de contornos en forma de cigarro. Cada contorno significa una dosis particular y todos los puntos interiores al contorno son puntos donde la dosis es mayor. Supondremos una velocidad del viento de 25 kilómetros por hora. En tal caso, la zona letal –la extensión circunscrita por la línea de contorno que denota una exposición de 400 rem en 24 horas– suma aproximadamente unos 1000 kilómetros cuadrados. El número de muertos en la zona letal dependerá estrechamente de la densidad de población. En los Estados Unidos, la densidad de población varía de 40.000 habitantes por kilómetro cuadrado en las áreas metropolitanas durante las horas comerciales a menos de dos por kilómetro cuadrado. Por ello, la radiación de la detonación de una simple cabeza nuclear puede matar varios centenares o varios millones de personas. El número total no sólo dependerá del punto de explosión, sino también de la hora del día, las condiciones climatológicas, la eficacia de cualquier aviso previamente notificado y de los medios de protección contra la radiación disponibles.

Quienes escapen a la muerte en la zona letal no podrían regresar a la zona por largo período, debido a la contaminación del territorio por partículas radiactivas. Los supervivientes tendrían que esperar hasta que los efectos de la desintegración radiactiva y la filtración de los contaminantes en el suelo a través de la lluvia y la nieve redujeran la radiactividad a un nivel aceptable. Para



vaporizara el núcleo de un reactor de un gigawatt. En este caso, la radiactividad tanto del arma nuclear como del reactor se dispersaría por todo el territorio afectado. En todas las hipótesis el viento que prevalece es del oeste, a 25 kilómetros por hora. La pluma de residuos podría desplazarse en un cierto número de direcciones (*círculos grises*). Las dosis se dan en rem por año. Un rem designa la cantidad de radiación que deposita 100 erg en un gramo de tejido. La radiación natural de fondo, a nivel del mar, es aproximadamente de 0,075 rem por año. La exposición de una población a 2 rem en el período de un año podría aumentar la incidencia de cáncer a largo plazo. Una exposición a 50 rem en un año puede causar enfermedades por radiación. Los puntos grises indican los lugares donde se han construido o están en proyecto reactores a menos de 40 kilómetros de una ciudad de más de 100.000 habitantes.

una dosis máxima aceptable de 2 rem por año, quedarían inservibles unos 3000 kilómetros cuadrados de tierra a lo largo de un año. Zonas mayores se verían afectadas por períodos inferiores. Los trastornos para la sociedad serían inmensos. Piénsese, por ejemplo, que más de 50.000 kilómetros cuadrados podrían resultar inhabitables a lo largo de todo un mes. Ni que decir tiene que ello supondría el abandono de su hogar por parte de muchos cientos de miles de personas.

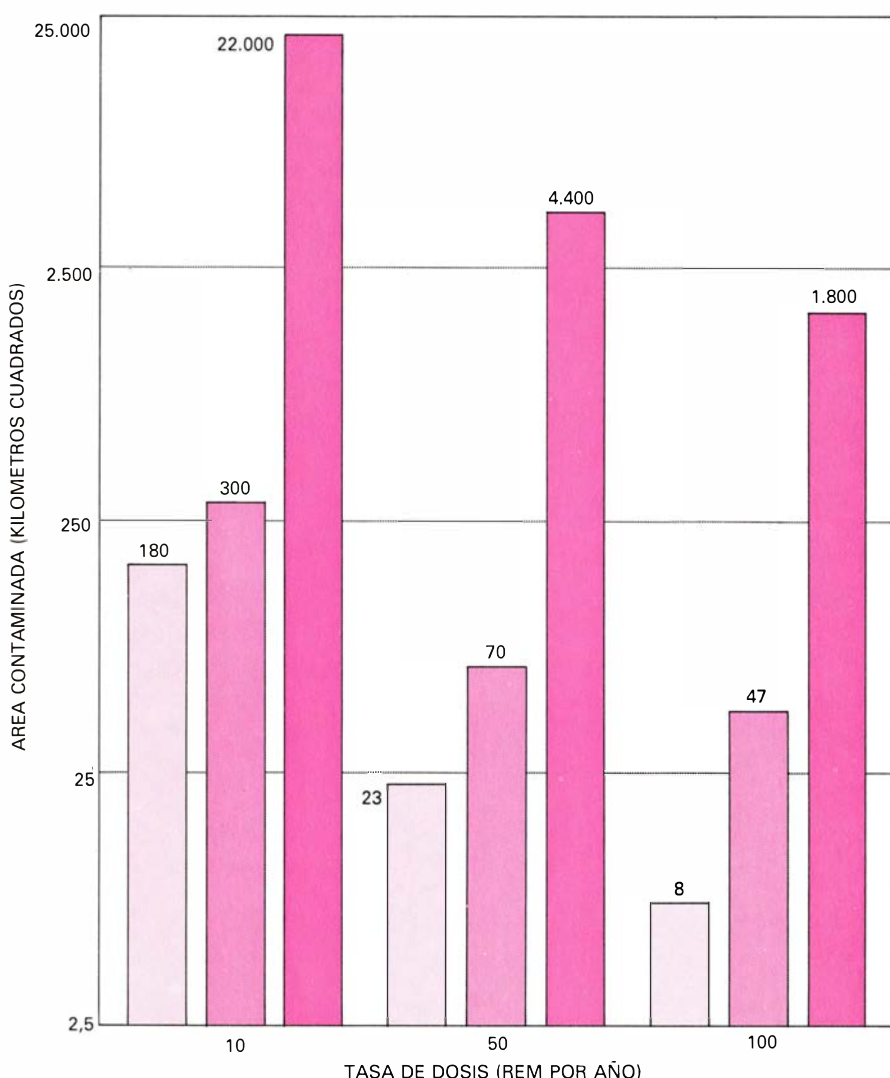
En un ataque en que explosionaran varias cabezas nucleares, la radiactividad acumulada podría impedir, casi con absoluta certeza, que la población superviviente volviera a sus puestos de trabajo y a las explotaciones agrarias que se hubieran librado de la destrucción por el impacto. Aun cuando los supervivientes pretendieran asentarse en zonas donde se hallaran expuestos a dosis superiores a 2 rem por año, no podrían ocupar grandes extensiones de territorio. Cada bomba de un megatón crearía una zona de casi 4000 kilómetros cuadrados donde la dosis de radiación se mantendría, a lo largo de un mes entero, por encima de los 50 rem anuales.

A diferencia de las armas nucleares, los reactores no pueden explotar. El reactor nuclear libera energía mediante fisión nuclear, pero incluso en un reactor completamente fuera de control la tasa de liberación de energía muestra una lentitud del orden de 10^{12} veces inferior a la tasa que detenta un arma nuclear. Además, la energía liberada en el reactor se absorbe inicialmente por la masa del núcleo del reactor, centenares de veces superior a la masa del arma atómica. La temperatura del núcleo se eleva paulatinamente incluso en un reactor descontrolado.

Si la temperatura del núcleo del reactor alcanzara un valor muy alto, se fundirían los elementos combustibles y el núcleo podría romperse antes de que una reacción en cadena generase cantidades de energía suficientes para provocar una explosión. Una rotura de la vasija de contención del reactor podría ocasionar una liberación de radiactividad. En un accidente verosímil, la pérdida total de refrigerante en las barras de combustible del núcleo ocasionaría el sobrecalentamiento y fusión de las barras. El material fundido establecería contacto con el agua y la explosión del vapor producido rompería la vasija de contención, produciéndose acto seguido una liberación del material radiactivo. En otro tipo de accidente imaginable, el sobrecalentamiento del núcleo generaría hidrógeno u otros gases inflamables que se mezclarían con el oxígeno atmosférico, pudiendo entonces producirse la ignición y explotar. También aquí se rompería la vasija de contención y habría un escape de radiactividad.

Al objeto de comparar los daños producidos por una liberación de radiactividad de un reactor con la liberación causada por un arma nuclear, examinaremos las consecuencias de estos accidentes, los peores posibles, que implican la rotura de la vasija de contención. Debe señalarse que la probabilidad de que suceda uno de estos acontecimientos se ha calculado en varios órdenes de magnitud menor que la probabilidad de un accidente más leve, tal como el que ocurrió en la Three Mile Island Nuclear Generating Station, cerca de Harrisburg, Pennsylvania, en marzo de 1979.

La cantidad de material radiactivo que escaparía de un reactor y la composición del mismo dependerían de la naturaleza exacta del accidente y del tiempo transcurrido desde que se recargó el reactor por última vez. La dispersión de la radiactividad dependería de la forma de la pluma de las emisiones liberadas por el accidente y de las condiciones meteorológicas locales. Sur-



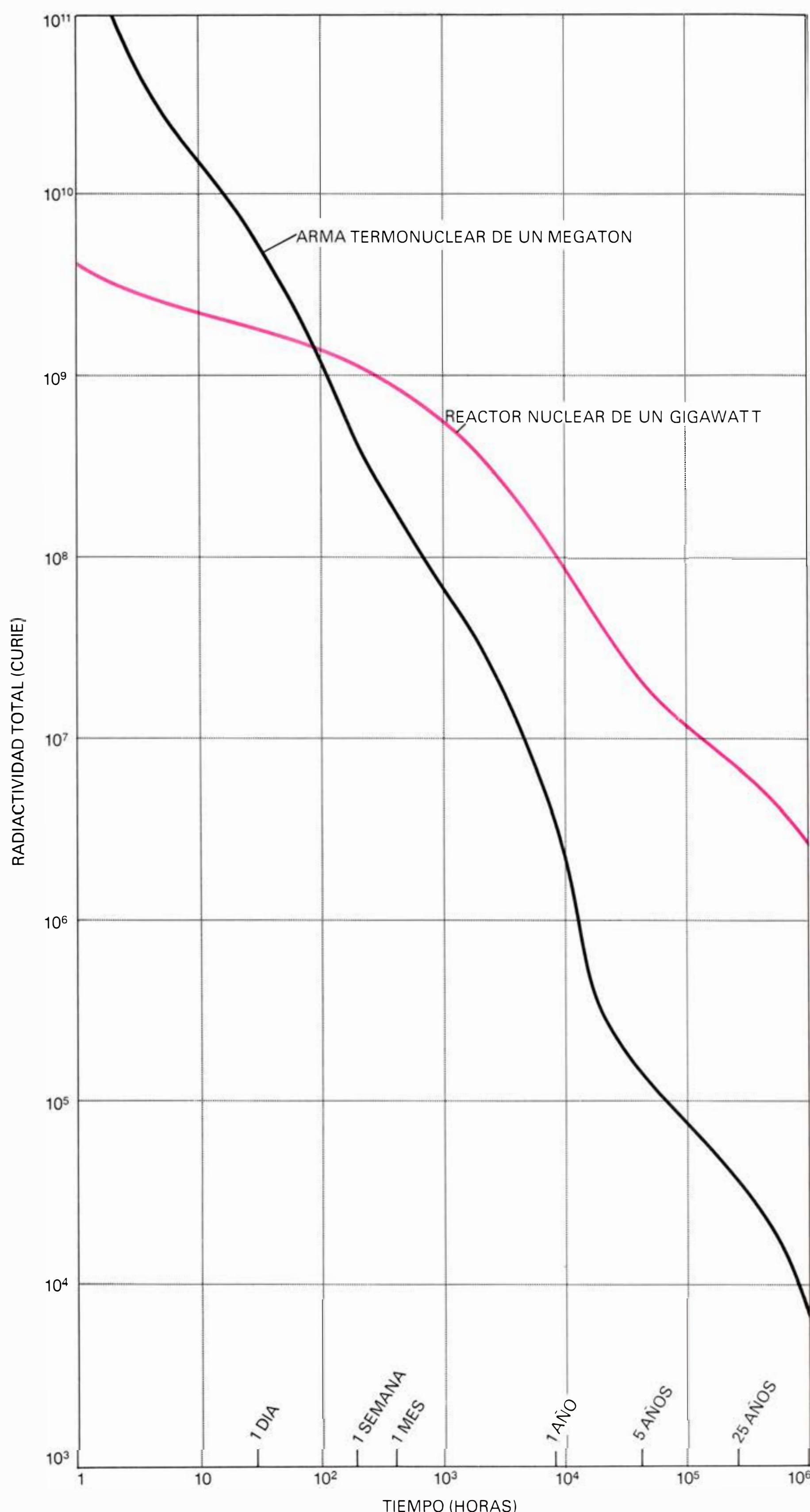
PROHIBICION DEL USO del terreno a los supervivientes de una liberación de radiactividad en función de la dosis de radiación que estén dispuestos (o sean obligados) a absorber. Una tasa de dosis de muy pocos rem por año sería inaceptable para un accidente en tiempo de paz, mientras que los supervivientes de un ataque nuclear podrían soportar bastante más. Las barras muestran la cantidad de terreno que debe permanecer vedado durante un año, si la dosis máxima aceptable es de 10 rem por año (*izquierda*), 50 rem por año (*centro*), o 100 rem por año (*derecha*). De nuevo, se han considerado tres causas posibles de contaminación radiactiva: un accidente grave en un reactor (*color claro*), la detonación de un arma nuclear en la superficie del terreno (*color intermedio*) y la detonación de un arma termonuclear sobre un reactor (*color intenso*). Si se considera inaceptable una dosis superior a 10 rem por año, la extensión de terreno que debe permanecer cerrada durante un año después del ataque es de 22.000 kilómetros cuadrados.

gen dos conclusiones de tipo general. En primer lugar, la velocidad de liberación de radiactividad por un arma termonuclear es, inicialmente, mucho mayor que la liberación de radiactividad en el accidente de un reactor nuclear. Sin embargo, la radiactividad del arma nuclear tiene una proporción mucho mayor de isótopos radiactivos de vida corta. En segundo lugar, en el accidente de un reactor nuclear se libera, comparativamente hablando, muy poco calor. En virtud de ello, la pluma de contaminación permanece a poca altura y deposita su radiactividad bastante de prisa, lo que tiende a limitar la extensión de la superficie contaminada. Resumiendo: aunque la extensión contaminada por el accidente de un reactor es mucho menor, el terreno permanece contaminado a lo largo de un período mayor.

Consideraremos un reactor nuclear de un gigawatt (1000 megawatt) en el que se reemplace anualmente un tercio del combustible. Supongamos que una explosión rompa la vasija de contención y libere a la atmósfera un tercio de los núcleos radiactivos contenidos en el reactor. Una hora después del escape, habrá que estimar la radiactividad del material liberado en unos 1500 millones de curie. La detonación de una cabeza termonuclear de un megatón liberaría una radiactividad 1000 veces superior. El accidente de la central de Three Mile Island liberó 100 millones de veces menos radiactividad. (Se liberaron 17 curie de iodo radiactivo.)

Supongamos, de nuevo, que la velocidad del viento es de 25 kilómetros por hora. El hecho crucial de las consecuencias de un accidente de este tipo es que la zona contaminada es bastante pequeña. Sin embargo, la exposición a la radiación permanece cerca del nivel de 2 rem por año para los habitantes de la región contaminada, salvo una zona muy pequeña. Concretamente, la dosis se mantiene en 2 rem por año durante un mes en un área de unos 4500 kilómetros cuadrados. (La cifra comparable para la detonación de una cabeza nuclear de un megatón es de más de 50.000 kilómetros cuadrados.) La zona letal, donde la dosis alcanza un nivel de 400 rem por día, es inferior a 2,5 kilómetros cuadrados. (La zona letal para la detonación de una cabeza nuclear de un megatón se cifra en los 1000 kilómetros cuadrados.)

Las menores dosis y la menor extensión del área contaminada en el caso de



DESINTEGRACION DE LA RADIACTIVIDAD liberada por la detonación de un arma nuclear. Difiere de la desintegración de la radiactividad liberada en el accidente de un reactor, porque los inventarios respectivos de núcleos radiactivos tienen distintas proporciones de varios isótopos. Pasada una hora, la radiactividad liberada por la detonación de un arma termonuclear de un megatón es 1000 veces mayor que la radiactividad que escaparía en el peor accidente imaginable en un reactor en tiempos de paz. La radiactividad liberada en el accidente de un reactor tarda más en disminuir. La unidad de radiactividad es el curie. Un curie corresponde a $3,7 \times 10^{10}$ emisiones por segundo de varias formas de radiación.

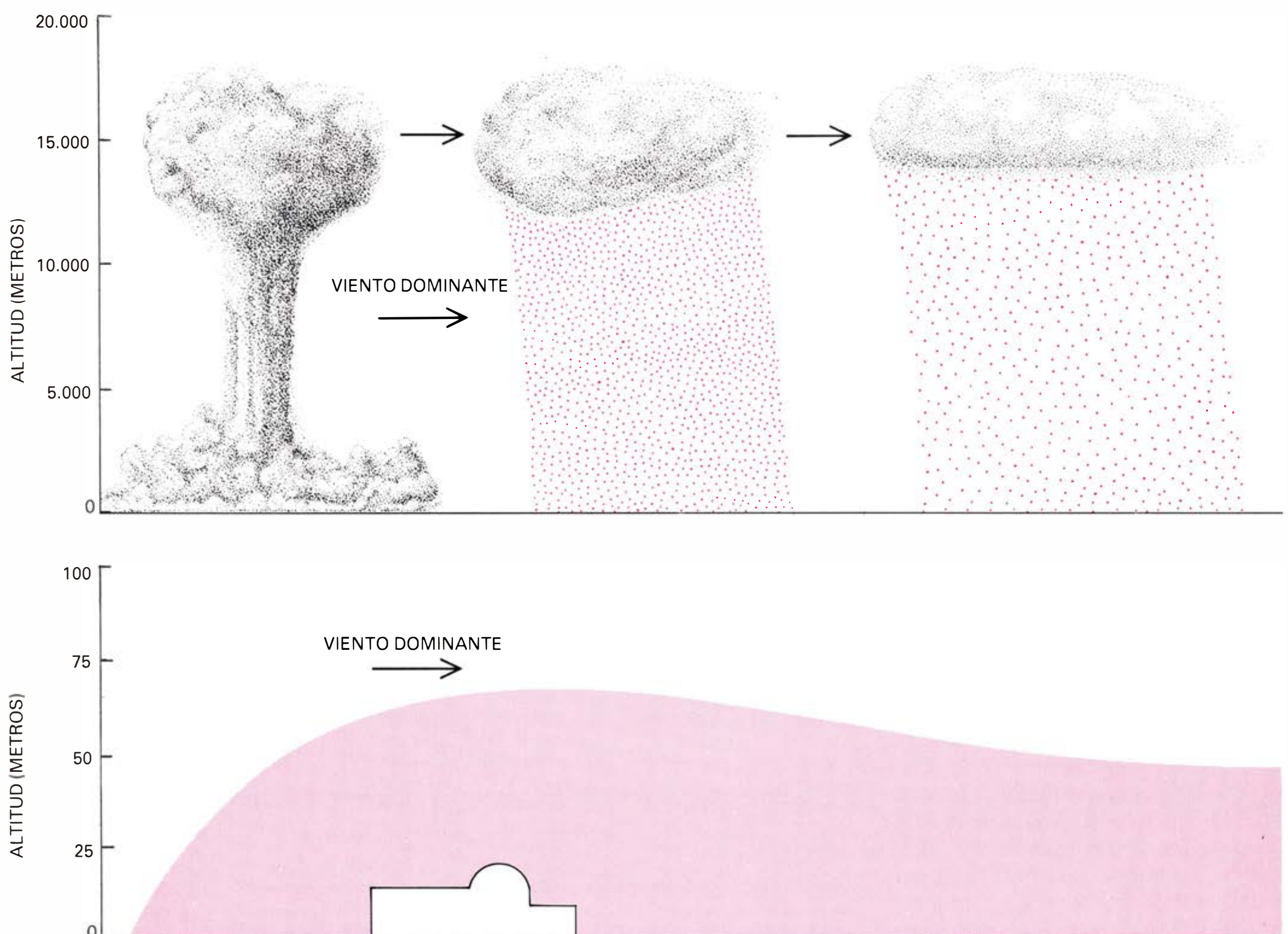
un accidente de un reactor sugiere que podría evacuarse la población antes de que inhalara cantidades sustanciales de polvo radiactivo (que es el principal peligro después del accidente en el reactor). También parece posible la descontaminación del territorio. En el caso de armas nucleares, la descontaminación del terreno resultaría imposible por ser mucho más elevados los depósitos de radiactividad.

Al comparar los efectos destructores del accidente de un reactor con los causados por la detonación de un ingenio nuclear, resulta útil considerar brevemente los efectos de destrucción inmediata producidos por ambos sucesos. La fusión del núcleo de un reactor no causa daños significativos por explosión o calor. Por el contrario, un arma nuclear acarrea la devastación inmediata en 8 o 16 kilómetros a la redonda del punto de explosión. En consecuencia, y con toda probabilidad, la detonación de una cabeza nuclear destruirá o dañará gravemente las instalaciones médicas y de emergencia. Por ello, es razonable concluir que si se expusiera una comunidad a una dosis peligrosa de radiación producida por un ataque nuclear y otra población fuese expuesta a la misma dosis liberada por el accidente de un reactor nuclear, aquélla tendría muchos menos supervivientes que ésta. La razón habría que buscarla en la grave alteración de los servicios que necesitarían las víctimas de una exposición a los efectos radiactivos.

No es fácil evaluar con certeza la probabilidad de los dos sucesos que hemos considerado. Sin embargo, parece ser que la opinión de los expertos en asuntos de la defensa y los especialistas en energía nuclear es que la probabilidad de la detonación de un arma nuclear en algún lugar del mundo en los próximos diez años resulta bastante más elevada que la probabilidad de una fusión de carácter catastrófico en un reactor nuclear. Entre las razones que podrían citarse en apoyo de este criterio están: el crecimiento constante de los arsenales de armas atómicas de varias naciones, la truculencia que caracteriza las relaciones entre los Estados Unidos y la Unión Soviética y los esfuerzos de los estrategas militares en desplazar la política estratégica de sus respectivos países hacia la preparación para la intervención en una guerra nuclear en vez de tratar de evitarla.

Pasemos a ponderar la radiactividad que se liberaría en el caso de que un arma termonuclear de un megatón detonara sobre un reactor nuclear de un gigawatt. Supondremos que todo el material radiactivo del núcleo del reactor se vaporiza completamente por efecto de la explosión. La radiactividad del reactor se combinaría con la radiactividad producida por el arma nuclear; ambas se elevarían con la bola de fuego y volverían a la tierra en la forma característica de lluvia radiactiva de una simple explosión nuclear.

Dado que la tasa de radiactividad en el reactor es inicialmente mucho menor que la tasa de radiactividad producida por la detonación de un arma nuclear, el esquema de contaminación en la pri-



PLUMA DE LOS RESIDUOS producidos por la detonación de un arma nuclear; difiere también de la pluma causada por el accidente de un reactor. La detonación de una cabeza termonuclear de un megatón de potencia en la superficie del terreno (*dibujo superior*) crea un empuje hacia arriba que eleva los residuos de la explosión a una altitud de unos 18.000 metros. Los residuos arrastrados en la dirección del viento vuelven a la tierra en forma de lluvia radiactiva. Por contra, la explosión no nuclear que rompe la vasija de contención de un reactor (*dibujo inferior*) tiene poca energía, de modo que la radiactividad liberada no alcanza gran altitud. La ausencia casi total de pluma reduce la diseminación de la contaminación producida por el viento predominante.

mera semana no diferiría, de forma apreciable, del esquema diseñado para el caso del arma nuclear solamente. Sin embargo, dado que la radiactividad procedente del reactor posee una vida relativamente larga, el tiempo que un área determinada permanecerá contaminada es significativamente mayor. En esencia, los residuos del arma nuclear contribuirían a un alto nivel de contaminación en el intervalo inmediato a la explosión, y los residuos del reactor contribuirían a la radiactividad de larga duración. La zona letal afectada por la detonación del arma sería de más de 1300 kilómetros cuadrados, un tercio mayor que la zona letal creada por la mera detonación de un arma nuclear. El área donde la dosis acumulada permanecería de 2 rem al año, durante un mes, sería de 165.000 kilómetros cuadrados, o sea, tres veces más extensa. El área donde las dosis permanecerían por encima de 2 rem por año, durante todo un año, sería de 64.000 kilómetros cuadrados, o sea, 20 veces superior. Un área de 460 kilómetros cuadrados continuaría durante más de un siglo exponiendo a sus habitantes a una dosis mínima de 2 rem por año. Dicha región constituiría un monumento permanente a la catástrofe.

Queda patente, pues, que la vaporización de los núcleos de los reactores nucleares con armas atómicas constituye un método eficaz para destruir y devastar grandes zonas de una nación. Sin ningún género de dudas, esperando las condiciones climáticas adecuadas, un país beligerante dispuesto a realizarlo, o en situación desesperada, podría arrasar un fracción substancial de la capacidad industrial de su antagonista con una sola arma termonuclear. Por ejemplo, un ataque a un reactor en el valle de los ríos Rhin y Neckar podría volver inhóspita un tercio de Alemania Occidental, un área de casi 250.000 kilómetros cuadrados, a lo largo de un mes o más, aun cuando dosis acumuladas de radiactividad muy superiores a 2 rem por año fuesen aceptables para los supervivientes. La única condición es que el ataque se realizase cuando los vientos dominantes procedieran del sureste.

Al reflexionar en torno a una devastación de este tipo, conviene tener presente que, en Europa central, donde la densidad de población es alta y se cultiva intensivamente la tierra, las centrales nucleares pudieran hallarse no muy lejos de instalaciones militares. Por ello, la probabilidad de que un arma nuclear destinada a un objetivo militar destruya fortuitamente un reactor nuclear cercano a la misma no es despreciable. Importa recordar también que las piscinas de almacenamiento de los residuos radiactivos de los reactores están situadas junto al reactor que los produce. Los residuos radiactivos de una piscina de almacenamiento típica pueden alcanzar fácilmente una cantidad de radiactividad dos veces mayor que la del propio reactor. Por si fuera poco, las centrales nucleares suelen construirse en unidades de dos reactores distantes entre sí escasos centenares de metros. Si se consideran todas estas circunstancias, las dosis de radiación que seguirían a la detonación de un arma nuclear sobre una central compleja, como la descrita, pueden duplicar y hasta sextuplicar el valor de las descritas anteriormente.

No podemos encontrar ningún documento público de que los estrategas militares hayan considerado, en ninguno de los escenarios supuestos de una guerra nuclear, la vaporización, accidental o deliberada, del núcleo de un reactor nuclear en un ataque atómico. La mejor manera de reducir al mínimo la probabilidad de este riesgo es evitar todo tipo de guerra nuclear. Pasos útiles serían la negociación de un acuerdo internacional de no designar como objetivos militares las instalaciones nucleares y los esfuerzos por conseguir que las instalaciones militares no se ubiquen cerca de los reactores civiles.

Si hubiéramos de sacar una conclusión final del análisis expuesto, ésta sería que una sola arma nuclear, en caso de explotar, contaminaría una zona mucho mayor que el peor accidente que puede concebirse en un reactor nuclear. En vista de ello, parece estar un poco fuera de lugar la preocupación del público por los riesgos que presenta la generación de electricidad mediante reactores nucleares. Un accidente catastrófico en un reactor causaría, sin duda, alteraciones muy importantes en su inmediata vecindad. Probablemente, ocasionaría problemas médicos a largo plazo, e incluso la pérdida de algunas vidas. No obstante, el impacto del accidente podría mitigarse y moderarse, porque los servicios médicos, sociales y gubernamentales quedarían intactos y en pleno funcionamiento, dentro de la misma área contaminada inclusive. Además, los riesgos que conllevan los reactores pueden minimizarse por la aplicación inteligente de la tecnología. Pero el ataque nuclear difiere radicalmente. Y no olvidemos que, hoy, la guerra nuclear encierra en sí misma un peligro de muerte y sufrimientos en una escala sin precedentes en la historia de la humanidad.

Teoría unificada de las partículas elementales y las fuerzas

A un alcance de 10^{-29} centímetros, el mundo puede resultar muy simple, con sólo una clase de partículas elementales y una fuerza importante. De ser correcta la teoría unificada que se propone, toda la materia sería inestable

Howard Georgi

No puede existir nada más simple que una partícula elemental: trozo indivisible de materia, sin estructura interna y sin tamaño o forma detectables. Análoga simplicidad podríamos esperar en la teoría que describe estas partículas y las fuerzas a través de las que interaccionan. Cabría sospechar, cuando menos, que la estructura del mundo pudiera explicarse con un número mínimo de partículas y fuerzas. A la luz de este criterio de simplicidad hemos de considerar como éxito razonable una descripción de la naturaleza que se ha desarrollado en estos últimos años. La materia está formada por sólo dos clases de partículas elementales: los leptones, el electrón por ejemplo, y los quarks, que son los constituyentes del protón, del neutrón y de otras muchas partículas análogas. Cuatro fuerzas básicas interactúan entre las partículas elementales. La gravitación y el electromagnetismo, que desde hace tiempo nos son familiares en el mundo macroscópico; la fuerza débil y la fuerza fuerte, que sólo se observan en sucesos subnucleares. En principio, este elenco de partículas y fuerzas podría dar cuenta de toda la jerarquía observada en las estructuras materiales, desde los núcleos de los átomos hasta las estrellas y las galaxias.

Ya es todo un logro entender la naturaleza a ese nivel de detalle. Pero cabe imaginar cómo sería una teoría más sencilla todavía. No acaba de satisfacer que existan dos clases dispares de partículas elementales; lo ideal sería que sólo hubiera una. Avanzando en esa dirección, la existencia de cuatro fuerzas parece una complicación innecesaria; una fuerza podría explicar todas las interacciones de las partículas elementales. Dentro de ese marco, una nueva y ambiciosa teoría promete, al menos, una unificación parcial. En dicha teoría no se incluye la gravitación, la más débil, con mucho, de las fuerzas y que, quizá, difiera fundamentalmente de las otras. Pero si prescindimos de la gravitación, la teoría unifica todas las partículas elementales y todas las fuerzas.

Se dio el primer paso hacia la construcción de una teoría unificada al demostrarse que las interacciones débiles, fuertes y electromagnéticas podían describirse, en su integridad, por teorías de una misma clase. Aunque distintas, podía verse que las tres fuerzas actuaban según el mismo mecanismo. En el curso de este desarrollo se descubrió una profunda conexión entre las fuerzas débiles y el electromagnetismo, conexión que apuntaba hacia una síntesis mayor. La nueva teoría se ha erigido en el principal candidato para lograr la síntesis. Incorpora los leptones y los quarks en una sola familia y contiene un método para transformar una partícula de una clase en una de la otra. Al mismo tiempo, las interacciones débiles, fuertes y electromagnéticas constituyen aspectos distintos de una única fuerza fundamental. Con una sola clase de partículas y una fuerza (más la gravitación), la teoría unificada es un modelo de frugalidad.

Se sabe que leptones y quarks tienen propiedades muy distintas. ¿Cómo pueden englobarse, pues, en una sola familia? Las fuerzas débiles, fuertes y electromagnéticas difieren en intensidad, alcance y otras características, ¿cómo pueden derivarse de una sola fuerza? La teoría unificada no pretende ocultar las diferencias, sino que afirma que no son fundamentales. Las diferencias destacan por la razón principal de que el universo es ahora muy frío, de forma que las partículas tienen, en general, energías bajas. Si pudiéramos realizar los experimentos a energías sumamente altas, la unificación se revelaría en toda su simplicidad. Leptones y quarks se transformarían unos en otros con absoluta libertad, y las tres fuerzas mostrarían la misma intensidad.

La energía necesaria para contemplar la unificación de las partículas y de las fuerzas de esta forma contundente se estima en unos 10^{15} gigaelectronvolt, que se abrevia GeV. (Un GeV es la energía que adquiere un electrón al ser acelerado por una diferencia de potencial de 1000 millones de volt.) Esta energía excede las capacidades de los mayores aceleradores de partículas proyectados en un factor de 10 billones; resulta muy improbable que tal energía se alcance nunca en el laboratorio. De ahí que pudiera parecer que la teoría no se someterá nunca a comprobación. Nada menos cierto. De la teoría se desprenden unas consecuencias bien definidas a energías fácilmente accesibles.

En primer lugar, la teoría aporta una

SURGE LA SIMETRIA CUBICA cuando ciertas propiedades de las partículas elementales se representan gráficamente en tres dimensiones. Las partículas son miembros de las familias llamadas leptones y quarks. La posición de cada partícula en el plano horizontal viene determinada por tres clases de "carga de color". Los quarks llamados u aparecen en tres clases de colores y están en los vértices de un triángulo equilátero; los antiquarks $\bar{d}$ poseen los tres anticolores correspondientes y configuran un triángulo orientado de forma opuesta. Los leptones, representados aquí por el positrón (e^+) y el neutrino (ν), carecen de carga de color y se hallan en el centro del plano. Cuando cada partícula se desplaza verticalmente, por una distancia proporcional a su carga eléctrica, surge un cubo. El hecho de que esta distribución de partículas de origen a un sólido simple y simétrico sugiere alguna conexión entre los leptones y los quarks. Una conexión explicable a través de una teoría unificada que englobe todas las partículas elementales.

+1

e^+

$+2/3$

u u u

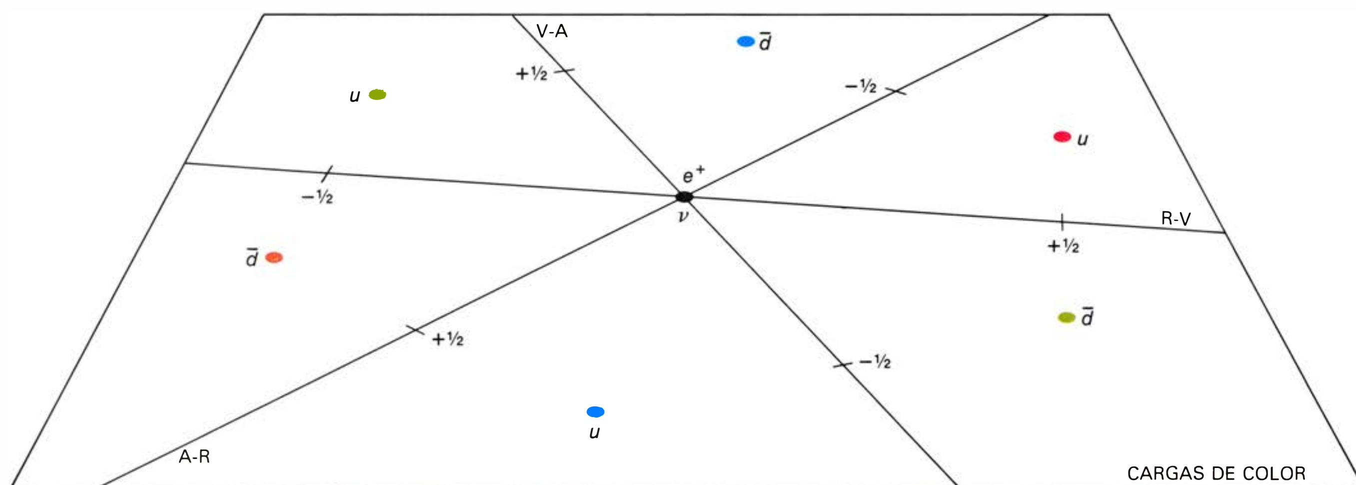
CARGA ELECTRICA

$\bar{d}$ $\bar{d}$ $\bar{d}$

$+1/3$

ν

0



justificación racional de varios hechos conocidos del mundo físico que durante mucho tiempo, y debido a su arbitrariedad, estuvieron rodeados del mayor misterio. Explica la cuantificación de la carga eléctrica: la observación de que ésta aparece siempre en forma de múltiplos discretos de una carga más pequeña y fundamental. Da un valor para las intensidades relativas de las tres fuerzas (medidas a las energías usuales en el laboratorio) que concuerda, razonablemente bien, con los resultados experimentales. Podría explicar por qué en el universo hay más materia que antimateria. Importa asimismo destacar que la teoría unificada predice nuevos fenómenos que no pueden deducirse de teorías anteriores. De esas predicciones vale resaltar la desintegración del protón, una partícula que se había considerado totalmente estable. Y si el protón puede desintegrarse, los mismos átomos serán inestables y, perecedera, toda la materia.

La teoría unificada no pretende suplantar las teorías establecidas de las fuerzas débiles, fuertes y electromagnéticas. Al contrario, las tres se engloban en una estructura mayor. Para explicar la naturaleza y el origen de la teoría unificada es mejor, por tanto, empezar por cada teoría particular, por las fuerzas que describen y por las partículas elementales sobre las que actúan dichas fuerzas.

Las diferencias que se manifiestan entre leptones y quarks son notables. Se conocen seis leptones, cuyo prototipo podría ser el electrón. Dotado de una masa pequeña, equivalente en unidades de energía a unos 500.000 electronvolt, tiene una unidad de carga eléctrica; por convención, la carga del electrón es negativa. Otros dos leptones, el muon y la partícula llamada tau, poseen la misma carga y parecen ser idénticos al electrón en todas sus propiedades, salvo en la masa. La del muon supera, en más de 200 veces, la masa del electrón; la del leptón tau, descubierto hace sólo cinco años, viene a multiplicar por 3500 veces la del electrón.

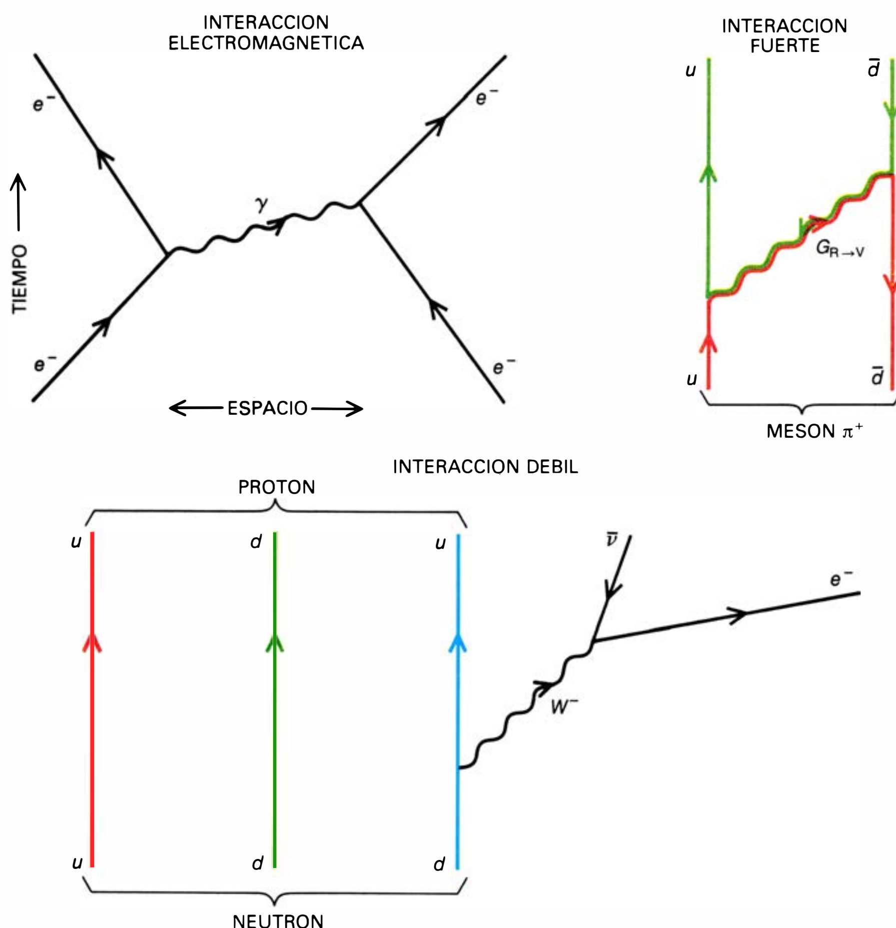
Los leptones restantes comprenden tres clases de neutrinos, eléctricamente neutros y cuya masa es muy pequeña (si es que tienen masa). Cada leptón cargado posee un neutrino asociado. Además, por cada uno de los seis leptones hay un antileptón, dotado de la misma masa pero con carga eléctrica opuesta. El antielectrón (o positrón), el antimuon y el antitau presentan todos, pues, carga +1. Los antineutrinos, al igual que los neutrinos, carecen de carga eléctrica.

Mientras que los leptones se encuentran como partículas libres, nadie ha podido examinar todavía ningún quark aislado. Los quarks se observan sólo como constituyentes de las partículas llamadas hadrones, una clase amplia y variada que abarca el protón, el neutrón, el mesón pi y más de otras 100 partículas conocidas.

Hay abundantes pruebas en favor de la existencia de cinco clases de quarks:

	LEPTONES		QUARKS					
TERCERA GENERACION	ν_τ	0	t	$+2/3$	t	$+2/3$	t	$+2/3$
	τ^-	-1	b	$-1/3$	b	$-1/3$	b	$-1/3$
SEGUNDA GENERACION	ν_μ	0	c	$+2/3$	c	$+2/3$	c	$+2/3$
	μ^-	-1	s	$-1/3$	s	$-1/3$	s	$-1/3$
PRIMERA GENERACION	ν_e	0	u	$+2/3$	u	$+2/3$	u	$+2/3$
	e^-	-1	d	$-1/3$	d	$-1/3$	d	$-1/3$

LEPTONES Y QUARKS difieren en un número de propiedades importantes, de ahí que se les haya generalmente clasificado en familias separadas. Una de las diferencias más drásticas está en la carga eléctrica, dada aquí para cada partícula: las cargas de los leptones son enteras, mientras que las cargas de los quarks son fraccionarias. Más aún, los leptones existen como partículas libres, en tanto que los quarks se encuentran sólo como constituyentes de partículas compuestas, llamadas hadrones. Se acostumbra dividir los leptones y los quarks en tres generaciones; sólo las partículas de la primera generación tienen un lugar en la estructura de la materia ordinaria. El quark *t* no se ha observado por vía experimental.



TRES FUERZAS de la naturaleza son las responsables de las interacciones entre partículas elementales. Cada una de estas interacciones puede describirse como el intercambio de una partícula "virtual", que es la portadora de la fuerza. En una interacción electromagnética las partículas con carga eléctrica intercambian un fotón (γ). Las interacciones fuertes son transmitidas por los gluones (G), que son intercambiados por partículas con carga de color. Partículas con carga débil pueden intercambiar un W^- (*representado aquí*), o W^+ o un Z^0. La carga de una antipartícula se indica por una flecha que apunta hacia el pasado.

se les denomina abajo ("down", d), arriba ("up", u), extraño ("strange", s), encanto ("charm", c) y fondo ("bottom", b). Aunque se ha predicho la existencia de una sexta clase de quark, el llamado cima o superior ("top", t), no se ha dado todavía con él. Las clases de quarks reciben también el apelativo de sabores. Los quarks poseen, además, otra propiedad: color. (Sabor y color son designaciones arbitrarias, que no guardan relación alguna con sensaciones gustativas o visuales.) Un quark de un sabor dado puede aparecer en tres colores: rojo, verde y azul. La propiedad del color establece una diferencia importante entre leptones y quarks. Los cinco o seis sabores de los quarks se corresponden, de una forma aproximativa, con las seis variedades de leptones, pero no existe entre los leptones el análogo al color de los quarks. La divergencia entre aquéllos y éstos comporta consecuencias observables. Las interacciones fuertes se deben a la interacción entre colores. Puesto que los leptones carecen de color, no sufrirán las interacciones fuertes.

Otra propiedad distintiva de los quarks es su carga eléctrica. Los quarks d, s y b tienen carga $-1/3$, en tanto que los quarks u, c y t poseen carga $+2/3$. Los antiquarks, que se denotan por $\bar{d}$, $\bar{u}$, etcétera, presentan valores opuestos de la carga eléctrica; así pues, la carga del antiquark $\bar{d}$ será $+1/3$ y, $-2/3$, la del antiquark $\bar{u}$. Los antiquarks tienen también colores opuestos; es decir, antirrojo, antiverde y antiazul.

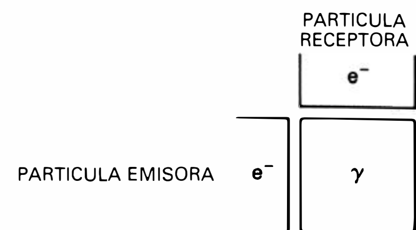
Para formar un hadrón, los quarks pueden combinarse de dos maneras: ligándose entre sí tres quarks, con un quark de cada color, o bien ligándose un quark de un color dado con un antiquark del anticolor correspondiente. Estas combinaciones se denominan blancas o sin color. Poseen, además, otra propiedad característica. En todas las combinaciones permitidas, se suman las cargas eléctricas fraccionarias de los quarks para dar una carga total entera; no hay otras combinaciones (excepto los múltiplos de las permitidas) que tengan esta propiedad. El protón está compuesto por los quarks uud, siendo su carga eléctrica total $+2/3 +2/3 -1/3$, es decir, $+1$. El neutrón consta de los quarks udd, con cargas de $+2/3 -1/3 -1/3$, dando una carga total nula. El mesón pi positivo está constituido por un quark u y un antiquark $\bar{d}$; las cargas de los componentes $+2/3$ y $+1/3$ dan una carga total de $+1$.

El hecho de que todos los átomos sean eléctricamente neutros implica que la carga del protón posea exactamente la misma magnitud que la del electrón, aunque, por supuesto, sus signos sean opuestos. Por idéntica razón, la carga del neutrón debe ser exactamente cero. De estas observaciones se deduce que las cargas de los quarks han de conmensurarse, de una manera cabal, con las de los leptones. Por ejemplo, la carga del quark d debe ser un tercio justo, y no sólo aproximado, de la del electrón. Esta relación precisa entre partículas que parecen ser independientes es otra propiedad que se diría que está ahí de un modo azaroso y que, sin embargo, requiere ser explicada en el marco de una teoría unificada.

Se acostumbra clasificar los leptones y los quarks en tres generaciones. Cada generación está formada por un leptón cargado, su neutrino asociado y dos quarks, uno de carga $-1/3$ y el otro con carga $+2/3$. En la primera generación se encuadra el electrón, el neutrino de tipo electrónico, el quark d y el quark u. Puesto que los quarks tienen tres colores habrá ocho partículas por generación. Todos los átomos y toda la materia ordinaria puede formarse a partir de estas ocho partículas; las otras generaciones se observan, de modo casi exclusivo, en experimentos de laboratorio, con partículas aceleradas. Aunque en la teoría unificada las tres generaciones se describen de forma independiente, el proceso esencial es el mismo. Me limitaré, por tanto, a examinar sólo la primera generación.

De las tres fuerzas que consideraré, el electromagnetismo fue la primera en recibir un tratamiento teórico preciso; precisión que ninguna otra teoría ha superado. La teoría que estudia esa fuerza es la electrodinámica cuántica o QED. Se desarrolló a lo largo de unos veinticinco años, que culminaron a comienzos de la década de 1950. Ha servido de modelo para las teorías que abordan las otras fuerzas.

La idea de fuerza se halla en íntima relación con la de carga. Por carga eléctrica se entiende la propiedad atribuida a una partícula que responde a las fuerzas electromagnéticas, y la cantidad de carga determina la respuesta. Cuando dos partículas cargadas se aproximan mutuamente, se establece una atracción o una repulsión cuya magnitud es directamente proporcional al producto de las cargas. La fuerza es también inversamente proporcional al cuadrado de la distancia entre las cargas. Estas dos reglas constituyen la ley de Coulomb de la fuerza eléctrica. Importa destacar que si una de las partículas tiene carga cero, no hay atracción ni repulsión; tales partículas neutras no se muestran directamente susceptibles en presencia de la fuerza electromagnética.



DESCRIPCION del electromagnetismo por una simetría que se representa por $U(1)$, expresión tomada de la teoría matemática de grupos. $U(1)$ es el grupo de transformaciones que pueden ser transmitidas por un objeto único o en una matriz uno por uno. En su aplicación al electromagnetismo, la simetría $U(1)$ implica que la fuerza electromagnética no puede cambiar la identidad de la partícula. La matriz uno por uno está ocupada por el fotón, que sólo puede transformar un electrón en otro.

¿Cuán fuerte es la interacción electromagnética entre partículas cargadas? Para unas partículas determinadas, la contestación dependerá de las cargas y de su distancia de separación, pero la ley de Coulomb puede darnos una respuesta general. Supongamos que la fuerza entre las dos partículas se multiplica por el cuadrado de la distancia entre ambas: el producto mide la intensidad de la interacción electromagnética, que, siendo independiente de la separación de las partículas, de-

	ROJO	VERDE	AZUL
ROJO	$G_1 + G_2$	$G_{R \to V}$	$G_{R \to A}$
VERDE	$G_{V \to R}$	$G_1 + G_2$	$G_{V \to A}$
AZUL	$G_{A \to R}$	$G_{A \to V}$	$G_1 + G_2$

LA FUERZA FUERTE viene descrita por una teoría con una simetría $SU(3)$, en la que los acoplamientos de los gluones a los quarks pueden representarse mediante una matriz tres por tres. Cualquier color de los quarks de la columna del extremo izquierdo de la matriz se puede transformar en cualquiera de los colores de la fila superior de la matriz; la transición viene mediada por el gluon especificado en la intersección entre la fila y la columna. Un quark rojo determinado puede emitir un gluon $G_{R \to A}$ y transformarse en un quark azul. Dos gluones no alteran el color, sino que median transformaciones tales como rojo pasa a rojo.

pende del sistema de unidades en que se exprese la separación. Dividiendo por la velocidad de la luz y por la constante de Planck (dos cantidades que aparecen en la estructura de una teoría mecánico-cuántica relativista del mundo) da un resultado que nada tiene que ver con las unidades; es decir, resulta un número puro o sin dimensiones; tiene el mismo valor cuando las medidas se hacen en gramos, centímetros y segundos o cuando se trabaja en toneladas, pies y años, si la velocidad de la luz y la constante de Planck se expresan en las unidades usadas en las medidas.

La intensidad de una interacción electromagnética dada depende del tamaño de las cargas que intervienen. Así, la interacción sería cuatro veces mayor si ambas cargas se duplicaran. Por hallarse la carga eléctrica cuantificada, la interacción de dos protones o de dos electrones desempeña un papel especial. Todas las partículas que han sido aisladas (es decir, todas las partículas a excepción de los quarks) tienen cargas que son múltiplos enteros de la carga del protón; así, la interacción protón-protón constituirá una medida de la intensidad mínima de la interacción electromagnética. Esta cantidad se denomina la constante de acoplamiento electromagnética, y es una medida absoluta de la intensidad de la interacción. Determinaciones experimentales de la constante de acoplamiento dan un valor de alrededor de 1/137. Al tratarse de un valor menor que 1, la interacción electromagnética es débil.

Hemos de puntualizar que la cuantificación de la carga ni la exige ni la predice la electrodinámica cuántica; no es más que un hecho experimental. La teoría se mostraría con igual coherencia si existieran partículas observables con cargas fraccionarias o, incluso, con cantidades irracionales de carga, tales como pi o la raíz cuadrada de 2.

En electrodinámica cuántica, la interacción entre dos partículas cargadas, dos electrones por ejemplo, está relacionada con el intercambio de una tercera partícula. La partícula intermediaria es un fotón: un cuanto de radiación electromagnética. El fotón, partícula sin masa, carece de carga eléctrica propia y se mueve (por definición) con la velocidad de la luz. La descripción de la fuerza electromagnética como intercambio de fotones evita la idea incómoda de acción a distancia. La interacción queda confinada a dos sucesos puntuales: la emisión y la absorción del fotón. Pero la descripción introduce, al propio tiempo, otro problema nada trivial: el intercambio de un fotón parece violar las leyes de la naturaleza que exigen que la energía y el momento deben conservarse.

Podemos ilustrar la violación aparente imaginando dos electrones estacionarios, separados por una cierta distancia. Puesto que podría medirse una fuerza entre los electrones, habrá que suponer que los fotones se están intercambiando. De ordinario, cuando se emite un fotón, éste se lleva parte de la energía y momento de la partícula emisora; de manera similar, cuando un fotón es absorbido, se añade al momento y a la energía de la partícula absorbente. La cantidad total de energía y momento del sistema se conserva, pues. En la situación considerada aquí, sin embargo, la partícula emisora se mantiene estacionaria y, por tanto, su energía y su momento no pueden cambiar; y lo mismo vale decir para la partícula absorbente. El fotón intercambiado tiene propiedades especiales, distintas de las de los fotones, que forman la luz del sol o las ondas de radio. En razón de esa diferencia del fotón intercambiado se le llama fotón virtual.

La explicación de estas propiedades peculiares del fotón virtual está en el principio de indeterminación introducido en la mecánica cuántica por Werner Heisenberg. El principio de indeterminación no invalida las leyes de conservación de la energía y el momento, pero permite que no se note una violación de estas leyes si se rectifica con suficiente rapidez. Los electrones estacionarios tiene idéntica energía y momento antes de emitir el fotón virtual y después de que éste haya sido absorbido; las leyes de conservación parecen violarse sólo durante el breve paso del fotón. El principio de indeterminación establece que tal violación manifiesta puede tolerarse si no dura demasiado tiempo o no tiene un alcance excesivo.

¿Qué significa aquí demasiado tiempo y demasiado grande? Las contestaciones variarán según sea la magnitud

CARGAS DE COLOR

		R-V	V-A	A-R	
QUARKS	ROJO	+½	0	−½	= 0
	VERDE	−½	+½	0	= 0
	AZUL	0	−½	+½	= 0
		=0	=0	=0	
GLUONES	G_1	0	0	0	= 0
	G_2	0	0	0	= 0
	$G_{R \to V}$	+1	−½	−½	= 0
	$G_{V \to R}$	−1	+½	+½	= 0
	$G_{V \to A}$	−½	+1	−½	= 0
	$G_{A \to V}$	+½	−1	+½	= 0
	$G_{R \to A}$	+½	+½	−1	= 0
	$G_{A \to R}$	−½	−½	+1	= 0
		=0	=0	=0	

PODEMOS IDENTIFICAR LAS CARGAS DE COLOR de los quarks y los gluones como rojo menos verde (R − V), verde menos azul (V −A) y azul menos rojo (A − R). Cada uno de los colores de los quarks, rojo, verde y azul, viene definido por una combinación de las tres cargas. Es significativo que las cargas que contribuyen a cada color se suman a cero. Esto implica que las tres cargas no son totalmente independientes; en realidad, bastan dos para identificar el color de una partícula. (Aquí mantenemos las tres cargas por razones de mayor claridad.) Para un triplete de quarks integrado por un quark de cada color, la suma de los valores de cada una de las cargas es también nula. Seis de los gluones tienen cargas de color, con los valores precisos para convertir un quark de un color en otro de otro color. La distribución de cargas en el triplete de quarks y la presencia de cargas en los gluones exigen la cuantificación de la carga de color: los únicos valores posibles de la carga de color son múltiplos enteros de 1/2 unidad.

de la violación que ocurra: cuanto mayor sea la violación de energía y momento causada por la emisión de un fotón virtual, antes deberá reabsorberse el fotón. Un fotón virtual de alta energía puede sobrevivir sólo brevemente, mientras que otro de baja energía gozará de un largo período de gracia antes de que los libros de balance se deban ajustar. Para ser explícito, el producto de la violación de la conservación de la energía y la vida media del fotón no pueden superar la constante de Planck. La energía mínima que puede ostentar cualquier partícula es el equivalente energético de la masa en reposo de las partículas y, por tanto, el alcance máximo de una partícula virtual dependerá inversamente de su masa. El alcance de la fuerza electromagnética parece ser infinito y, por tanto, la masa en reposo del fotón deberá ser cabalmente nula.

La presencia de partículas virtuales complica mucho la estructura del universo. Debido a ellas, el vacío no es un mero espacio sin nada. Un fotón virtual puede aparecer espontáneamente en cualquier instante y desaparecer de nuevo en el tiempo permitido por el principio de indeterminación. De igual forma, pueden crearse otras partículas virtuales, sin excluir las cargadas eléctricamente; la única restricción que ha de cumplirse es que las partículas con una carga eléctrica deben aparecer y desaparecer en pares formados por la partícula y la antipartícula. Este proceso tiene profundas consecuencias en la teoría del electromagnetismo.

COMBINACION DE ANTIQUARKS

	R-V	V-A	A-R
VERDE	−½	+½	0
+ AZUL	0	−½	+½
ANTIRROJO	−½	0	+½
ROJO	+½	0	−½
+ AZUL	0	−½	+½
ANTIVERDE	+½	−½	0
ROJO	+½	0	−½
+ VERDE	−½	+½	0
ANTIAZUL	0	+½	−½

COMBINACION DE GLUONES

	R-V	V-A	A-R
ROJO	+½	0	−½
+ ANTIVERDE	+½	−½	0
$G_{R\to V}$	+1	−½	−½
ROJO	+½	0	−½
ROJO	+½	0	−½
+ AZUL	0	−½	+½
$G_{R\to V}$	+1	−½	−½

UN PROCEDIMIENTO que predice correctamente las propiedades de color de los antiquarks y de los gluones es suponer que son combinaciones de los quarks que forman el triplete fundamental de color. Cualquier antiquark (que debe tener cargas de color opuestas a las del quark correspondiente) puede "formarse" añadiendo las cargas de color de dos quarks. Cualquier gluon puede formarse a partir de los colores de un quark y un antiquark y, puesto que el antiquark también puede descomponerse, a partir de los colores de tres quarks. Así se explican todas las combinaciones posibles. Este procedimiento es sólo formal; los antiquarks y los gluones no deben considerarse como físicamente compuestos de quarks.

Consideremos qué sucede cuando un electrón real está inmerso en una nube de fotones virtuales y pares electrón-positrón virtuales. Los fotones apenas si se dejan sentir, pero las partículas virtuales cargadas se polarizan: cargas virtuales negativas son repelidas por la carga real negativa, mientras que cargas virtuales positivas son atraídas por el electrón real. Resulta así que el electrón se encuentra rodeado, en su inmediata vecindad, por una nube de cargas positivas, que apantallan parte de la carga del electrón.

De este análisis se deduce que la carga "desnuda" del electrón es mucho mayor que la carga medida. En realidad, en la electrodinámica cuántica se supone que la carga desnuda es infinita. La carga medida constituye sólo el residuo finito que queda cuando, de la carga desnuda, se resta la carga apantallante. Si pudiera medirse la carga del electrón desde distancias extraordinariamente cercanas, hallaríamos que aumenta a medida que se fuera penetrando en el apantallamiento. Síguese de ello, pues, que la constante de acoplamiento del electromagnetismo no es en absoluto una constante, sino que varía con la distancia a la que se encuentran las partículas cargadas que interaccionan entre sí. La constante de acoplamiento aumenta (lo que significa que la interacción electromagnética se hace más fuerte) cuando se reduce el alcance. La constante de acoplamiento medida de, aproximadamente, 1/137 es la observada a distancias atómicas de unos 10^{-8} centímetros.

En el propio mundo efímero de las partículas virtuales hay una ley de conservación que no se viola nunca: la conservación de la carga eléctrica. Por ser neutro el fotón, la carga se conserva automáticamente en el intercambio de un fotón virtual: las cargas no se alteran. Más aún, cuando se crea o aniquila materia cargada, lo es siempre en pares de partículas y antipartículas, de forma que, después del suceso, las cargas son las mismas que antes.

La conservación de la carga eléctrica y el hecho de que el fotón no tenga masa están relacionados con un grupo de simetrías en el sistema matemático que describe la electrodinámica cuántica. El grupo de simetrías se designa por $U(1)$. Decimos así que la QED es una teoría $U(1)$. $U(1)$ es un término empleado en la teoría matemática de grupos. El 1 se refiere al hecho de que el fotón interacciona con una sola clase de partícula en un instante. El fotón nunca transforma una partícula de una clase en otra partícula de otra clase. Las interacciones fuertes y débiles son más complicadas en este aspecto y, más complejos, los grupos que las describen.

La teoría de las interacciones fuertes que hoy prevalece asume como su modelo directo la electrodinámica cuántica. La teoría se llama cromodinámica cuántica, o QCD; "cromo-" significa que la fuerza actúa no entre las cargas eléctricas, sino entre las cargas de color. Como en QED, la magnitud de la fuerza entre dos cargas es proporcional al producto de las cargas; las partículas que no tienen carga de color no están sujetas a estas fuerzas. Una constante de acoplamiento sin dimensiones define la intensidad intrínseca de la interacción. La constante de acoplamiento es mayor que la constante del electromagnetismo, como debe esperarse en una fuerza que se llama fuerte.

Aunque la QCD se construye sobre los mismos principios que la QED, se trata de una teoría más elaborada. La fuente principal de la nueva complejidad reside en la multiplicidad de cargas de color. Mientras que el electromagnetismo está asociado con sólo una clase de carga, las fuerzas fuertes

actúan sobre tres colores: rojo, verde y azul. Cada color representa una combinación de las cargas de color subyacentes.

Hay varias formas distintas de analizar las cargas de color. La forma en que lo haré parte de la suposición de que hay tres clases de carga de color. Llamaré a las cargas rojo menos verde (R − V), verde menos azul (V − A) y azul menos rojo (A − R). Cada carga puede tener un valor +1/2, −1/2, o cero; cada color de un quark viene caracterizado por una combinación particular de estos colores. Un quark es rojo si tiene una carga R − V de +1/2, una carga V − A de 0 y una carga A − R de −1/2. Un quark verde tiene las siguientes cargas: rojo menos verde (R − V) = −1/2, V − A = +1/2 y A − R = 0. En un quark azul, las tres cargas de color son, respectivamente R − V = 0, V − A = −1/2 y, por último, A − R = +1/2. Los anticolores asociados con los antiquarks se forman, simplemente, cambiando los signos de todas las cargas.

Pueden hacerse varias observaciones sobre esta distribución de cargas. En primer lugar, existen 27 combinaciones posibles de las tres cargas, cuando cada una de ellas puede tener cualquiera de los tres valores. Sin embargo, parece que en la naturaleza sólo existen quarks con las tres combinaciones que dan los colores rojo, verde y azul. En segundo lugar, este subconjunto de las posibles cargas de color es muy peculiar. Cada combinación observada es tal que la suma de las tres cargas de color es nula; las combinaciones observadas son las únicas que tienen esta propiedad. (En realidad, hay otra combinación con una carga de color total nula: la combinación en la que cada carga es cero. Pero la partícula que no posee ninguna carga de color no es un quark.)

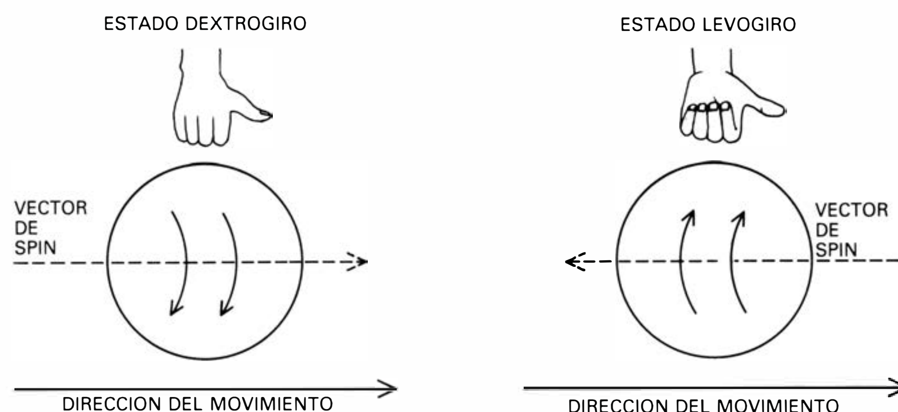
El hecho de que la suma de las tres cargas de color sea siempre nula indica que una de las tres cargas no es independiente de las otras dos. Si se conocen dos cualesquiera de las cargas, puede hallarse la tercera restando. De aquí podemos concluir que, en realidad, hay sólo dos variedades de carga de color, suficientes para especificar completamente los tres colores. Carece de importancia qué dos cargas se consideren fundamentales y cuál se elimine; aquí supondremos que las cargas R − V y V − A son las fundamentales, pero frecuentemente mantendré la carga A − R por claridad, aun cuando la información que suministra sea redundante.

Queda por describir una relación adicional entre las cargas. En un estado formado por un quark rojo, uno verde y uno azul, la cantidad total de carga de cada color será, de nuevo, cero. En otras palabras, combinando los tres colores se origina un estado neutro de color; análogamente a cómo de la combinación de electrón y protón se crea un estado (el átomo de hidrógeno) que es neutro con relación a la carga eléctrica. Así se forman los hadrones neutros de color, el protón, por ejemplo. Un sistema sin color puede también crearse combinando un color con el anticolor correspondiente; por ser opuestas, las cargas de color se cancelan exactamente. La otra fórmula para hacer un hadrón consiste en combinar un color con su anticolor; ocurre así con el mesón pi. Si exceptuamos los múltiplos de estas combinaciones (tales como un sistema de seis quarks, que abarque dos quarks de cada color), no hay otra manera de combinar los quarks coloreados de forma que todas las cargas de color tengan suma nula.

Podemos comparar el mecanismo por el que se transmiten las interacciones fuertes de color con el mecanismo correspondiente del electromagnetismo: la interacción entre dos partículas cargadas se describe como el intercambio de una tercera partícula. Y, una vez más, la cromodinámica cuántica se muestra como una teoría más elaborada. En tanto que la QED tiene un único fotón sin masa, QCD posee ocho partículas sin masa, llamadas gluones. Más aún: el fotón carece de carga eléctrica, pero algunos gluones llevan carga de color. La presencia de partículas portadoras cargadas altera, de manera fundamentalmente, el carácter de la fuerza en cuestión.

Por estar cargados a su vez, los gluones pueden alterar los colores de los quarks y no sólo transmitir las fuerzas fuertes. Por contra, la emisión o absorción de fotones nunca puede alterar la carga eléctrica de una partícula. Hay nueve transiciones posibles entre los colores de los quarks, definidos por una matriz tres por tres. Verbigracia: un quark rojo se puede transformar en un quark rojo (la transformación identidad), en un quark verde o en un quark azul. Las tres transformaciones identidad (rojo pasa a rojo, verde pasa a verde y azul pasa a azul) constituyen los elementos diagonales de la matriz. Es evidente que los gluones responsables de las transiciones identidad no pueden tener cargas de color o alterarían los colores de los quarks. Podría parecer que deberían existir tres gluones neutros de color, un gluon neutro por cada transformación identidad. Como hay sólo dos cargas de color independiente, que son necesarias para especificar los tres colores de los quarks, existen sólo dos gluones neutros de color. Las designaré por G_1 y G_2.

Las seis restantes transiciones entre los colores de los quarks implican cambios de color. Cada una de ellas está asociada con su propio gluon, y cada gluon posee una carga de color. Usaré un subíndice para describir los gluones con carga de color. Por ejemplo, un gluon rojo-a-verde, o $G_{R \to V}$, puede ser emitido por un quark rojo que, co-



HELICIDAD DE UNA PARTICULA, determinada por la orientación de su momento angular intrínseco de spin. Cuando el vector que define el eje de spin es paralelo a la dirección del movimiento de la partícula, se dice que la partícula es dextrógira; designación que obedece al hecho siguiente: cuando los dedos de la mano derecha se doblan de la forma en que gira la partícula, el pulgar indica la dirección del movimiento. Cuando el vector de spin es antiparalelo a la trayectoria, el pulgar de la mano izquierda da la dirección del movimiento y, por la misma razón anterior, la partícula se llama levógira. El electromagnetismo y la fuerza fuerte son indiferentes a la helicidad, pero ésta tiene una influencia importante en las interacciones débiles. Una partícula con masa puede cambiar su helicidad; no así una partícula sin masa.

mo consecuencia, se transforma en uno verde.

Las cargas de color que portan los gluones pueden deducirse de la necesidad de conservar dichas cargas de color. Consideremos el proceso en el que un quark cambia de rojo a verde por emisión de un gluon $G_{R \to V}$. En el curso de la transición, la carga R − V del quark cambia de +1/2 a −1/2; si debe permanecer constante la cantidad total de carga, el gluon habrá de tener, por tanto, una carga R −V de +1. Del mismo modo, la carga V − A del quark cambia de 0 a +1/2 y, por tanto, el gluon debe portar una carga V − A de −1/2. La carga A − R del quark pasa de −1/2 a 0, lo que implica que el gluon tiene una carga A − R de −1/2. Las cargas de color del gluon serán, respectivamente, +1, −1/2 y −1/2. El quark que media la transformación inversa, de verde a rojo, debe tener cargas de la misma magnitud y de signo opuesto.

La presencia de cargas de color en los gluones comporta una nueva consecuencia: automáticamente, asegura que la carga de color esté cuantificada. En el electromagnetismo, un fotón podría, en principio, ser emitido o absorbido por una partícula con cualquier carga eléctrica. Las partículas con carga de color pueden interaccionar intercambiando gluones, sólo si las cargas están separadas por intervalos que sean múltiplos de 1/2. También puede demostrarse que las cargas de color del sistema han de ser simétricas alrededor de cero, esto es, la suma de todas las cargas positivas y la suma de todas las cargas negativas debe ser igual en valor absoluto.

	CARGA DEBIL	CARGA $U(1)$	CARGA ELECTRICA	PARTICULAS	TRANSICIONES
DOBLETES	+½	−½	0	ν_{LEVO}	W^+ ↓ ↑ W^-
	−½		−1	e^-_{LEVO}	
	+½	+1/6	+⅔	u_{LEVO}	W^+ ↓ ↑ W^-
	−½		−⅓	d_{LEVO}	
	+½	+½	+1	e^+_{DEXTRO}	W^+ ↓ ↑ W^-
	−½		0	$\bar{\nu}_{DEXTRO}$	
	+½	−1/6	+⅓	$\bar{d}_{DEXTRO}$	W^+ ↓ ↑ W^-
	−½		−⅔	$\bar{u}_{DEXTRO}$	
SINGLETES	0	−1	−1	e^-_{DEXTRO}	
	0	+1	+1	e^+_{LEVO}	
	0	+⅔	+⅔	u_{DEXTRO}	
	0	−⅔	−⅔	$\bar{u}_{LEVO}$	
	0	−⅓	−⅓	d_{DEXTRO}	
	0	+⅓	+⅓	$\bar{d}_{LEVO}$	

LA CARGA DEBIL, además de depender de la helicidad de una partícula, guarda una relación curiosa con la carga eléctrica. Las partículas levógiras y las antipartículas dextrógiras forman dobletes en las interacciones débiles; se les asignan cargas débiles de más o menos 1/2. Los W^+ y W^-, que median las fuerzas débiles, transforman un miembro de un doblete en el otro miembro. Las partículas dextrógiras y las antipartículas levógiras permanecen en singletes y carecen de carga débil, de suerte que no hay transiciones débiles entre ellas. La carga eléctrica de cada partícula es, invariablemente, igual a la suma de la carga débil y de otra cantidad llamada la carga $U(1)$, que es igual a la carga eléctrica media de las partículas en el singlete o en el doblete del que forma parte la partícula. Esta relación hallada entre las cargas indica que debe haber una conexión subyacente entre la fuerza débil y el electromagnetismo.

La cuantificación de la carga de color puede demostrarse de otra forma. Cualquier sistema de partículas dotadas de color puede "construirse" a partir del más simple de ellos: el triplete formado por un quark rojo, uno verde y uno azul. El triplete de los antiquarks puede formarse combinando los quarks en pares. No pretendo sugerir que un antiquark físico conste de dos quarks en un estado ligado. Sin embargo, todas las propiedades de color de los antiquarks quedan correctamente descritas por esta síntesis. Notemos que un quark rojo tiene cargas de color R − V, V − A y A − R de +1/2, 0 y −1/2, respectivamente. Un antiquark antirrojo debe tener las cargas opuestas: −1/2, 0 y +1/2. Estos son exactamente los valores hallados añadiendo las cargas de un quark verde (−1/2, +1/2 y 0) y de un quark azul (0, −1/2 y +1/2). De aquí que el antirrojo sea, de alguna manera, equivalente a la suma de verde y azul. De igual forma, el antiverde consta del rojo más el azul y el antiazul, del rojo más el verde. Esta notable correspondencia es una simple consecuencia de la forma en que las cargas de color están distribuidas en el triplete de quarks. Por ser nula la carga total del triplete, la suma de dos cargas cualesquiera debe ser igual a la carga restante cambiada de signo.

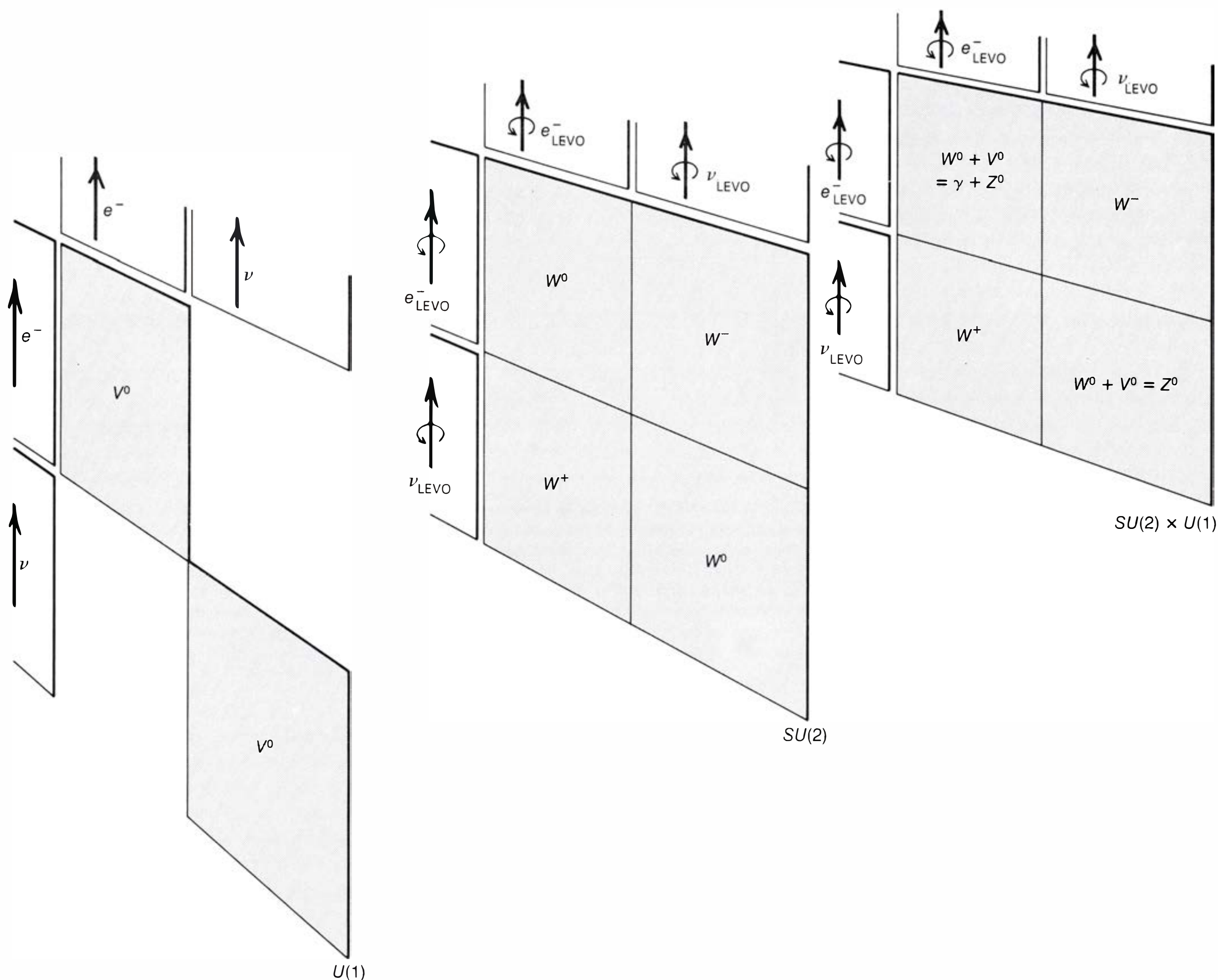
Los gluones pueden construirse de una forma similar, a partir de un quark y un antiquark, aunque no debe inferirse que físicamente un gluon sea un estado ligado de un quark y un antiquark. El gluon rojo-a-verde, con cargas +1, −1/2 y −1/2, puede imaginarse constituido por un quark rojo (+1/2, 0 y −1/2) y un antiquark antiverde (+1/2, −1/2 y 0). El antiquark antiverde puede descomponerse en un quark rojo y otro azul; por tanto, el gluon rojo-a-verde tiene las propiedades de color de dos quarks rojos y un quark azul.

Queda otra consecuencia derivada del hecho de que los gluones tengan cargas de color. Según se expuso antes, un electrón en el vacío está rodeado por una nube de fotones virtua-

les y pares electrón-positrón virtuales; las partículas virtuales cargadas se polarizan y apantallan una parte de la carga desnuda del electrón. Por el mismo mecanismo, un quark en un vacío queda rodeado por una nube de gluones virtuales y de pares quark-antiquark virtuales, si bien el resultado es totalmente distinto. La nube de quarks y antiquarks virtuales se polariza del modo acostumbrado, con los antiquarks apretujados cerca de la carga de color real y tendiendo a apantallarla. Los gluones virtuales muestran, por contra, el efecto opuesto. El color predominante de la carga de los gluones cerca del quark es el mismo que el de la carga del quark. Más aún, los gluones virtuales son más numerosos que los quarks virtuales, de forma que la influencia de los gluones es la más fuerte. El resultado es como si la carga del quark estuviera esparcida por el espacio, y la carga efectiva disminuyera a medida que nos acercáramos al quark.

En ausencia de cargas gluónicas, podría esperarse que la fuerza fuerte variara con la distancia, análogamente a cuanto acontece en el electromagnetismo. Puesto que los gluones no tienen masa, como el fotón, la fuerza tendría un alcance infinito, pero decrecería en intensidad a razón del cuadrado de la distancia. El hecho de que los gluones tengan cargas de color altera el carácter de la fuerza. Dado que la nube de gluones virtuales esparce la carga de color, la fuerza de color entre dos quarks no aumenta tan rápidamente como la fuerza electromagnética cuando se reduce la distancia entre las partículas. Y se sigue que la constante de acoplamiento de la QCD decrece cuando la distancia a la que se mide disminuye (a diferencia de la constante de acoplamiento de la QED, que aumenta a pequeña distancia). Se dice que los quarks son asintóticamente libres, lo que significa que la constante de acoplamiento de la QCD tiende hacia cero cuando la distancia se aproxima a cero.

La libertad asintótica fue descubierta por H. David Politzer, residente ahora en el Instituto de Tecnología de California, y por David Gross y Frank Wilczek, de la Universidad de Princeton. Sometida a comprobación, ha sido confirmada en múltiples experimentos que estudian la estructura quark de los hadrones a pequeñas distancias. Aun-



FUERZA DEBIL Y ELECTROMAGNETISMO pueden recibir un tratamiento conjunto a través de una teoría con una simetría que está representada por el producto de los dos grupos: $SU(2) \times U(1)$. La parte $SU(2)$ de la interacción induce todas las posibles transformaciones de dos objetos o de una matriz dos por dos. Los objetos son los miembros de los dobletes débiles, representados aquí por las componentes levógiras del electrón y el neutrino. Hay tres partículas $SU(2)$, sin masa: el W^+ y el W^- que convierten un electrón en un neutrino, y viceversa, y la W^0 que es la mediadora de las operaciones identidad (electrón va a electrón y neutrino a neutrino). La parte $U(1)$ de la interacción está asociada con otra partícula mediadora, la V^0, que sólo es capaz de efectuar las transformaciones identidad. En la teoría combinada $SU(2) \times U(1)$, las W^0 y V^0 se mezclan; las combinaciones observadas son el fotón (γ) y un portador de la fuerza débil, Z^0. La simetría $SU(2) \times U(1)$ se rompe espontáneamente, lo que determina que W^+, W^- y Z^0 tengan masas grandes.

que no está bien establecida la naturaleza de las interacciones fuertes entre los quarks a distancias mayores, parece que la fuerza no decrece a razón del cuadrado de la distancia, sino que permanece constante e independiente de ella. Sería, pues, necesaria una energía ilimitada para poder separar dos cargas de color, lo que explicaría por qué los quarks parecen confinados de un modo permanente dentro de los hadrones.

De la cromodinámica cuántica se dice que es una teoría $SU(3)$, apelativo este último que aparece también en teoría de grupos. El 3 se refiere a los tres colores que se transforman, unos en otros, a través de los gluones. La S indica que la suma de las cargas de color en cada familia $SU(3)$ es nula. Análogamente al $U(1)$ de la QED, el $SU(3)$ de la QCD describe un grupo de simetrías de la teoría que está asociada con la conservación de la carga de color y con el hecho de que los gluones carecen de masa.

El orden elevado de la simetría en la teoría de color $SU(3)$ no la puede poner de manifiesto una representación geométrica. Las tres cargas de color R − V, V − A y A − R pueden representarse por su posición respecto a tres ejes en el plano. Los ejes están simétricamente dispuestos a ángulos de 120 grados, los unos de los otros. Si se colocan los tres colores en razón de sus cargas de color que los componen, se encuentra que están en los vértices de un triángulo equilátero. Los anticolores, opuestos a los colores correspondientes, forman otro triángulo girado 180 grados con respecto al primero. Los dos triángulos superpuestos dibujan una estrella de David.

Si añadimos una tercera dimensión a la gráfica, obtendremos una pista de una simetría aún mayor en la teoría unificada. Supongamos que las cargas de color ya representadas sean las del quark u y las del antiquark $\bar{d}$. Añadamos ahora dos leptones: el neutrino de tipo electrónico y el positrón. Por carecer de carga de color, los leptones yacen en el origen de los tres ejes en el centro del plano. La tercera dimensión es la carga eléctrica: cada partícula debe ser desplazada verticalmente por una cantidad proporcional a su carga eléctrica. El neutrino permanece en su lugar, pero los tres antiquarks $\bar{d}$ son desplazados hacia arriba un tercio de una unidad, los quarks u se mueven hacia arriba dos tercios de una unidad y el positrón se mueve hacia arriba en una unidad. Si las escalas vertical y horizontal se eligen de forma adecuada, las ocho partículas definen los vértices de un cubo apoyado sobre un vértice. El mero hecho de que los quarks y los leptones puedan disponerse en la configuración de este sólido simple permite sospechar la existencia de alguna conexión profunda entre ellos.

Para examinar la última de las tres interacciones, la débil, es necesario introducir otra propiedad de las partículas elementales: el momento angular de spin. Se ha visto que leptones y quarks poseen todos la misma cantidad fija de momento angular, igual a 1/2, cuando se mide en unidades fundamentales. Se puede imaginar las partículas girando alrededor de un eje interno, igual que la tierra o una peonza, pero sin pérdida de energía. El momento angular se representa por un vector, o flecha, a lo largo del eje de giro.

Una partícula con medio cuanto de spin intrínseco puede tener sólo dos orientaciones posibles; en el caso más simple, cuando la partícula está en movimiento, el vector de spin puede hallarse en la misma dirección del movimiento o en dirección opuesta. Las dos orientaciones representan dos estados distinguibles de la partícula. Si el vector es paralelo a la dirección del movimiento, se dice que es dextrógiro: cuando los dedos de la mano derecha rodean la partícula en el mismo sentido que el spin, el pulgar indica la dirección del movimiento. Cuando el eje del spin está alineado de forma opuesta, el pulgar de la mano izquierda señala la dirección del movimiento y, por tanto, la partícula se dice que es levógira.

En general, el carácter levógiro o dextrógiro de una partícula puede cambiarse con sólo llevar la partícula al estado de reposo y acelerándola en la dirección opuesta sin perturbar el spin. Por tanto, muchas partículas tienen, necesariamente, tanto componentes levógiros como dextrógiros. Las excepciones son las partículas sin masa; ¿por qué son excepciones? Por la sencilla razón de que las partículas sin masa se mueven siempre con la velocidad de la luz y nunca pueden ponerse en reposo. Así pues, el carácter levógiro o dextrógiro de una partícula sin masa nunca puede cambiar. De entre quarks y leptones, las únicas partículas que pueden carecer de masa son los neutrinos. Por vía experimental sólo se han observado neutrinos levógiros y antineutrinos dextrógiros; se supone que no existen neutrinos dextrógiros ni antineutrinos levógiros.

La introducción de la helicidad (usaremos este nombre para indicar el carácter dextrógiro o levógiro) viene a doblar casi el número de partículas elementales distinguibles, un número que es ya bastante elevado. En la primera generación de partículas hay dos leptones (el electrón y el neutrino de tipo electrónico) y dos sabores de quarks (u y d). Los tres colores de los quarks dan un total de ocho partículas y, teniendo en cuenta las correspondientes antipartículas, se cuenta un total de 16. Si todas las partículas están dotadas de componente levógiro y dextrógiro, la introducción de la helicidad doblaría, de nuevo, el número de partículas. Al no haber neutrino dextrógiro ni antineutrino levógiro, el número total de partículas y antipartículas diferentes es de 30. Son estos 30 estados los que hay que acomodar en una teoría unificada.

Deben distinguirse los estados de diferente helicidad, porque las interacciones débiles actúan de modo diverso sobre los componentes levógiros y dextrógiros de una partícula. Igual que en las restantes fuerzas, la débil se encuentra asociada a una carga; la intensidad intrínseca de las interacciones débiles se puede definir mediante una constante de acoplamiento sin dimensiones. Sin embargo, la carga débil es poco usual en el sentido de que deviene asignada en función de la helicidad. Sólo las partículas levógiras y las antipartículas dextrógiras poseen carga débil; las partículas dextrógiras y las antipartículas levógiras son neutras en relación con la fuerza débil, y no participan en las interacciones débiles.

Por diferir la carga débil de un electrón levógiro de la de otro dextrógiro (por ejemplo), no puede conservarse la carga débil. El valor de ésta depende de la forma en que se mueva el electrón, valor que cambiará cuando lo haga el movimiento. La carga débil podría conservarse tan sólo si leptones y quarks fueran todos de masa nula, ya que en este caso ninguna de las partículas podría pararse e invertir el sentido del movimiento.

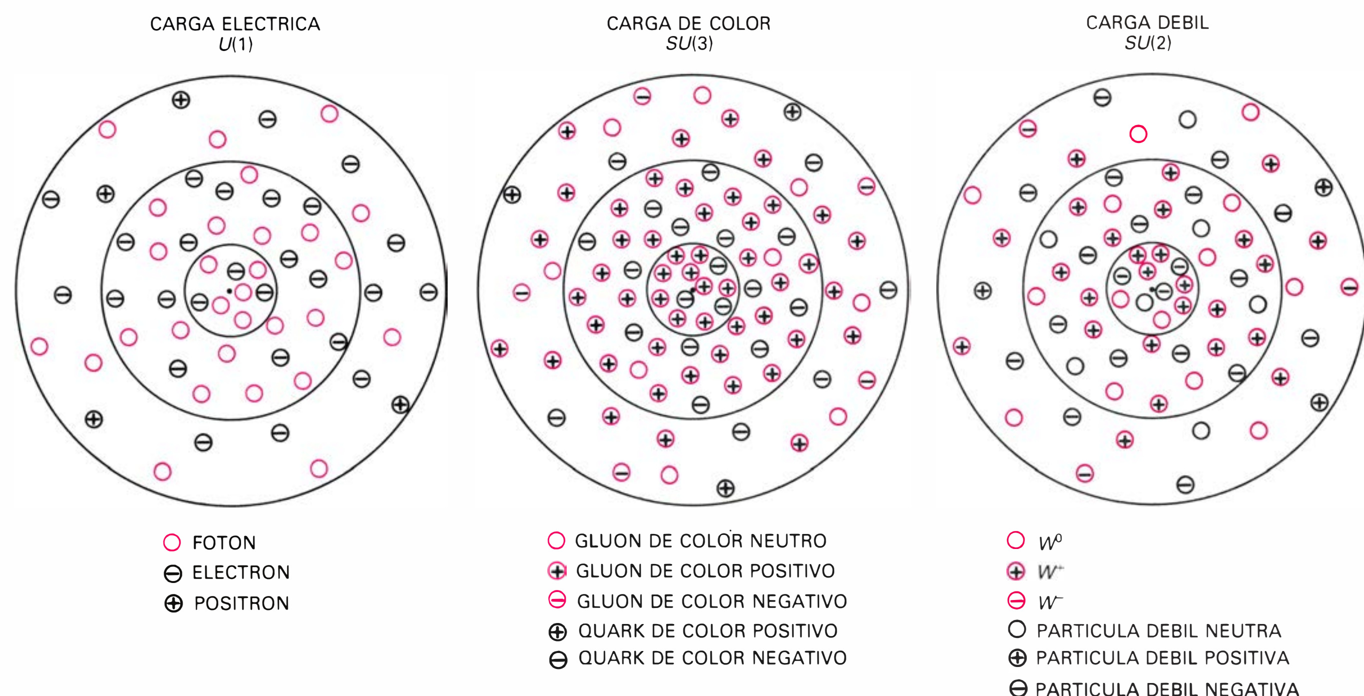
La fuerza débil actúa sobre dobletes de partículas. La teoría que describe es una teoría $SU(2)$, en la que los dos miembros del doble pueden transformarse entre sí. Por ejemplo, el neutrino levógiro y el electrón levógiro constituyen un doblete; se les asignan, respectivamente, cargas débiles de +1/2 y −1/2. El quark u levógiro y el quark d levógiro forman otro doblete (o tres dobletes si contamos cada color

por separado) y tienen también cargas débiles de +1/2 y −1/2, respectivamente. Las cuatro antipartículas dextrógiras forman los dobletes restantes: el positrón, el antineutrino electrónico, el antiquark $\overline{d}$ y el antiquark $\overline{u}$. Cada partícula dextrógira tiene una carga débil opuesta a la de la correspondiente partícula levógira. Se deben aún considerar las seis partículas restantes: las componentes dextrógiras del electrón, del quark *d* y del quark *u*, y las componentes levógiras del positrón, el antiquark $\overline{d}$ y el antiquark $\overline{u}$. No forman dobletes, sino que aparecen aisladas en forma de singletes, y tienen una carga débil nula.

Tres partículas asociadas con la simetría débil $SU(2)$ median las transiciones entre los miembros de cada doblete. Las partículas mediadoras son el W^+, con carga débil y con carga eléctrica +1; el W^-, con carga débil y eléctrica −1 y el W^0, que es neutro con respecto a las fuerzas débiles y electromagnéticas. El W^0, al igual que el fotón y los gluones G_1 y G_2, transmite una fuerza entre las partículas que llevan carga, pero no altera ninguna de sus propiedades. Por otra parte, los W^+ y W^- transforman los sabores de las partículas. Un electrón levógiro puede emitir un W^- y convertirse en un neutrino levógiro; en el proceso, la carga eléctrica cambia de −1 a 0 y la carga débil, de −1/2 a +1/2. El proceso débil más conocido es la desintegración beta nuclear, en la que un neutrón (cuya composición en quarks es *udd*) emite un electrón y un antineutrino y se convierte en un protón (*uud*). Analizado con más detalle, el proceso comienza cuando un quark *d* emite un W^- virtual y se convierte en un quark *u*; a continuación, el W^- se desintegra para dar un electrón y un antineutrino.

En sucesos como éstos se pueden percibir algunas relaciones exasperantes entre la fuerza débil y el electromagnetismo. En primer lugar, las partículas *W* de las fuerzas débiles llevan la misma cantidad de carga débil y de carga eléctrica. En segundo lugar, en la estructura de los singletes y dobletes débiles hay una curiosa relación fija entre la carga débil y la carga eléctrica. La carga eléctrica de una partícula es invariablemente igual a la suma de su carga débil y de la carga eléctrica media del singlete o doblete del que forma parte la partícula. Esta carga media la designaré como la carga $U(1)$. Para los singletes, la carga $U(1)$ es meramente la carga eléctrica de la partícula, y la regla no es mucho más que una tautología: dice que la carga eléctrica es igual a la carga eléctrica, ya que la carga débil de una partícula singlete es siempre nula. Sin embargo, para los dobletes débiles la relación de las cargas es más interesante. Se debe hacer notar que la relación permanece válida tanto para dobletes leptónicos, donde las cargas promediadas son enteros, como para los dobletes de quarks, que están hechos de cargas fraccionarias.

Al igual que las otras cargas, la carga $U(1)$ está asociada con una simetría. La simetría $U(1)$, extraída así de las interacciones débiles, tiene una partícula asociada, que llamaré V^0. Igual que la W^0 y el fotón, la V^0 no tiene carga eléctrica ni débil. En efecto, puesto que la carga débil, la $U(1)$ y la carga eléctrica están relacionadas, las tres partículas mediadoras deben también estarlo. Sucede que la W^0 y la V^0 no se observan en la naturaleza como estados puros y que aparecen sólo como mezclas. Una de estas mezclas de la W^0 y de la V^0 es el fotón; la otra combinación posible se identifica con la partícula llamada Z^0. Tanto el fotón como la Z^0 median interacciones en las que la partícula se transforma en sí misma; su identidad no cambia, pues. Difieren, sin embar-



UNA NUBE DE PARTICULAS VIRTUALES rodea una carga puntual central y altera su respuesta a una fuerza. Una carga eléctrica positiva está rodeada por fotones virtuales y por pares electrón-positrón virtuales. Los fotones tienen un efecto pequeño, pero las partículas virtuales cargadas están polarizadas, de forma que las cargas virtuales negativas se aglomeran alrededor de la carga positiva real, reduciendo su magnitud efectiva. Una carga positiva de color está rodeada de una nube de gluones virtuales y de pares virtuales quark-antiquark. Los quarks y los antiquarks se polarizan de forma análoga a como lo hacen los electrones y los positrones, pero los gluones actúan de forma distinta que los fotones. Mientras el fotón es eléctricamente neutro, algunos de los gluones tienen carga de color, predominantemente de la misma polaridad que la carga real. La carga de color queda, por tanto, esparcida por el espacio; la carga neta de cualquier volumen esférico se hace más pequeña cuando el radio de la esfera se reduce. El resultado final de estos efectos es que la interacción electromagnética crece a distancias pequeñas, en tanto que la interacción fuerte decrece. La interacción débil también tiene partículas portadoras cargadas, de manera que se hace más débil a distancias pequeñas, aunque ello ocurre en menor grado que las interacciones fuertes.

go, en que el fotón se acopla sólo a las partículas que tienen carga eléctrica, mientras que el Z^0 se acopla a aquellas que tienen carga débil, incluidos los neutrinos. Al originarse el fotón y las partículas W y Z de una misma teoría, la fuerza débil y el electromagnetismo quedan parcialmente unificados. La teoría se suele designar mediante el producto de los grupos que incorpora: $SU(2) \times U(1)$.

Por analogía con la simetría $U(1)$ de la electrodinámica cuántica y la $SU(3)$ de la cromodinámica cuántica, podría esperarse que la $SU(2)$ y la $U(1)$ de las interacciones débiles fueran simetrías exactas de la teoría. La carga débil se conservaría entonces exactamente y las partículas W y Z carecerían de masa y su alcance sería infinito. Pero según se dijo antes, la carga débil no se conserva invariablemente. Más aún, la fuerza débil observada tiene un alcance extraordinariamente corto, quizá de unos 10^{-15} centímetros. La razón estriba en que la W^+, la W^- y la Z^0 poseen una masa grande, próxima a los 100 GeV, es decir, 100 veces la masa del protón. ¿Qué ocurre con la simetría $SU(2) \times U(1)$ de las interacciones débiles y electromagnéticas en estas condiciones? Además, si el fotón y las partículas W y Z están muy relacionadas, ¿cómo puede una partícula quedar con masa nula y las otras adquirir una masa grande?

La contestación a tales preguntas constituía un paso esencial en la formulación de una teoría combinada de las interacciones débiles y electromagnéticas. La respuesta más aceptada hoy día afirma que la fuerza subyacente es, en efecto, simétrica y, en algún estado inicial hipotético, todos los mediadores de las interacciones débiles tienen masa nula. No es simétrico el vacío mecánico-cuántico, el medio donde actúa la fuerza. La estructura del vacío rompe espontáneamente la simetría $SU(2) \times U(1)$, dando masa a los tres mediadores de las fuerzas débiles, pero no al fotón. Puede observarse una pérdida análoga de simetría en un cristal de sal. Los iones del cristal no favorecen ninguna dirección del espacio frente a otra; los iones tienen simetría rotacional. Pero en la estructura de la red cristalina, la simetría se rompe y ciertas direcciones, tales como las paralelas a las líneas de la red, adquieren una especial relevancia.

La misma analogía puede iluminar otra característica de una simetría espontáneamente rota. El aspecto del

	CARGA ELECTRICA	CARGA DEBIL	CARGA R-V	CARGA V-A
d ROJO DEXTRO	−⅓	0	+½	0
d VERDE DEXTRO	−⅓	0	−½	+½
d AZUL DEXTRO	−⅓	0	0	−½
e^+ DEXTRO	+1	+½	0	0
$\bar{\nu}$ DEXTRO	0	−½	0	0
	= 0	= 0	= 0	= 0

FAMILIA COMPLETA de las partículas elementales, donde se integran tanto leptones como quarks. La familia contiene cinco miembros: tres quarks dextrógiros (los tres colores del quark *d*) y dos antileptones dextrógiros (el positrón y el antineutrino). A cada partícula se le asignan valores de la carga eléctrica, de la carga débil y de las dos cargas de color, que se identifican aquí como rojo menos verde y verde menos azul. La tercera carga de color se omite por ser redundante. Para cada clase de carga, la suma de los valores asignados a todas las partículas es cero. Más aún, en una teoría unificada, las partículas que llevan a cabo las transformaciones dentro de la familia portan cada una de las clases de carga. Las cargas de las partículas están, pues, cuantificadas. La teoría unificada está asociada con el grupo de simetría *SU*(5).

	d ROJO DEXTRO	d VERDE DEXTRO	d AZUL DEXTRO	e^+ DEXTRO	$\bar{\nu}$ DEXTRO
d ROJO DEXTRO	$G_1 + G_2 + \gamma + Z^0$	$G_{R \to V}$	$G_{R \to A}$	$X^{\text{ROJO}}_{-4/3}$	$X^{\text{ROJO}}_{-1/3}$
d VERDE DEXTRO	$G_{V \to R}$	$G_1 + G_2 + \gamma + Z^0$	$G_{V \to A}$	$X^{\text{VERDE}}_{-4/3}$	$X^{\text{VERDE}}_{-1/3}$
d AZUL DEXTRO	$G_{A \to R}$	$G_{A \to V}$	$G_1 + G_2 + \gamma + Z^0$	$X^{\text{AZUL}}_{-4/3}$	$X^{\text{AZUL}}_{-1/3}$
e^+ DEXTRO	$X^{\overline{\text{ROJO}}}_{+4/3}$	$X^{\overline{\text{VERDE}}}_{+4/3}$	$X^{\overline{\text{AZUL}}}_{+4/3}$	$\gamma + Z^0$	W^+
$\bar{\nu}$ DEXTRO	$X^{\overline{\text{ROJO}}}_{+1/3}$	$X^{\overline{\text{VERDE}}}_{+1/3}$	$X^{\overline{\text{AZUL}}}_{+1/3}$	W^-	Z^0

LA SIMETRIA *SU*(5) abarca todas las transiciones posibles entre partículas en la familia integrada de cinco. Las simetrías de las fuerzas individuales quedan incorporadas en la teoría *SU*(5) como subgrupos: la simetría *SU*(3) de las interacciones fuertes está incluida en la matriz tres por tres de la parte superior izquierda y la simetría *SU*(2) de la fuerza débil aparece en la matriz dos por dos de la parte inferior derecha. La simetría *U*(1), asociada con el Z^0, se presenta en los acoplamientos del fotón (γ) y el Z^0, a lo largo de la diagonal. La teoría *SU*(5) postula 12 nuevas partículas mediadoras, indicadas por *X*, que median las transformaciones de un quark en un leptón y de un leptón en un quark. La interconversión de quarks y leptones es sólo posible en una teoría unificada. En 1973 propuse ya una teoría basada en *SU*(5).

		CLASE DE CARGA			
		ELECTRICA	DEBIL	R-V	V-A
	$d^{\text{ROJO}}_{\text{DEXTRO}}$	−1/3	0	+1/2	0
+	e^{+}_{DEXTRO}	+1	+1/2	0	0
	$u^{\text{ROJO}}_{\text{LEVO}}$	+2/3	+1/2	+1/2	0
	$d^{\text{VERDE}}_{\text{DEXTRO}}$	−1/3	0	−1/2	+1/2
+	e^{+}_{DEXTRO}	+1	+1/2	0	0
	$u^{\text{VERDE}}_{\text{LEVO}}$	+2/3	+1/2	−1/2	+1/2
	$d^{\text{AZUL}}_{\text{DEXTRO}}$	−1/3	0	0	−1/2
+	e^{+}_{DEXTRO}	+1	+1/2	0	0
	$u^{\text{AZUL}}_{\text{LEVO}}$	+2/3	+1/2	0	−1/2
	$d^{\text{ROJO}}_{\text{DEXTRO}}$	−1/3	0	+1/2	0
+	$\bar{\nu}_{\text{DEXTRO}}$	0	−1/2	0	0
	$d^{\text{ROJO}}_{\text{LEVO}}$	−1/3	−1/2	+1/2	0
	$d^{\text{VERDE}}_{\text{DEXTRO}}$	−1/3	0	−1/2	+1/2
+	$\bar{\nu}_{\text{DEXTRO}}$	0	−1/2	0	0
	$d^{\text{VERDE}}_{\text{LEVO}}$	−1/3	−1/2	−1/2	+1/2
	$d^{\text{AZUL}}_{\text{DEXTRO}}$	−1/3	0	0	−1/2
+	$\bar{\nu}_{\text{DEXTRO}}$	0	−1/2	0	0
	$d^{\text{AZUL}}_{\text{LEVO}}$	−1/3	−1/2	0	−1/2
	e^{+}_{DEXTRO}	+1	+1/2	0	0
+	$\bar{\nu}_{\text{DEXTRO}}$	0	−1/2	0	0
	e^{+}_{LEVO}	+1	0	0	0
	$d^{\text{VERDE}}_{\text{DEXTRO}}$	−1/3	0	−1/2	+1/2
+	$d^{\text{AZUL}}_{\text{DEXTRO}}$	−1/3	0	0	−1/2
	$\bar{u}^{\overline{\text{ROJO}}}_{\text{LEVO}}$	−2/3	0	−1/2	0
	$d^{\text{ROJO}}_{\text{DEXTRO}}$	−1/3	0	+1/2	0
+	$d^{\text{AZUL}}_{\text{DEXTRO}}$	−1/3	0	0	−1/2
	$\bar{u}^{\overline{\text{VERDE}}}_{\text{LEVO}}$	−2/3	0	+1/2	−1/2
	$d^{\text{ROJO}}_{\text{DEXTRO}}$	−1/3	0	+1/2	0
+	$d^{\text{VERDE}}_{\text{DEXTRO}}$	−1/3	0	−1/2	+1/2
	$\bar{u}^{\overline{\text{AZUL}}}_{\text{LEVO}}$	−2/3	0	0	+1/2

SE OBTIENE UNA FAMILIA de 10 partículas formando todos los pares posibles de los cinco estados que constituyen la familia más simple de *SU*(5). Todos los miembros de la familia de cinco son dextrógiros, pero cuando se combinan en pares dan origen a estados levógiros. Como en el proceso de construir antiquarks a partir de pares de quarks, el método no debería interpretarse como una realidad física; una partícula levógira no es, en realidad, un estado ligado de dos partículas dextrógiras. Sin embargo, este procedimiento de construcción da, de forma correcta, todas las cargas de los estados levógiros. Con dos familias adicionales del mismo tamaño y de estructura similar, todas las partículas de la primera generación tienen un lugar en la teoría y no queda ningún lugar vacío.

mundo depende de la escala a la que se analice. En un cristal se aprecian hasta tres escalas de distancias. A distancias mucho menores de 10^{-8} centímetros (el tamaño de un átomo) se ve la estructura interna del átomo totalmente simétrica y que no viene afectada por la organización del cristal. A distancias de unos 10^{-8} centímetros, cobran importancia las fuerzas responsables de la cohesión de los átomos en el cristal, y los fenómenos observados son muy complejos. A distancias mucho mayores que 10^{-8} centímetros, se manifiesta con nitidez la geometría del cristal, y la simetría rotacional de los átomos está visiblemente rota.

En relación con la rotura espontánea de la simetría $SU(2) \times U(1)$ puede definirse una jerarquía similar de escalas de distancias, pero la distancia crítica es menor: alrededor de 10^{-16} centímetros. A distancias mucho menores aparece la simetría completa. A esa distancia tan corta, las partículas *W* y *Z*, que tienen masa, se intercambian con la misma facilidad que se aprecia en los fotones, que carecen de masa, y, por tanto, las fuerzas débiles y electromagnéticas quedan unificadas. Otra forma de expresar la misma idea es indicar una relación entre la distancia y la energía. De acuerdo con el principio de indeterminación, la energía necesaria para poder estudiar detalles de un cierto tamaño es inversamente proporcional al tamaño en cuestión. Un experimento que examinara la estructura de una partícula a un nivel menor que 10^{-16} centímetros tendría que llevarse a cabo con una energía superior a los 100 GeV. A esa energía, las partículas *W* y *Z* pueden crearse libremente, y la diferencia de masa entre ellas y el fotón es despreciable; comparadas con la energía del experimento, todas las partículas mediadoras son ligeras.

A una distancia de alrededor de 10^{-16} centímetros, empiezan a hacer

acto de presencia los complicados fenómenos responsables de la rotura de la simetría $SU(2) \times U(1)$. Aunque todavía se observan las partículas W y Z, son muy distintas del fotón, pues sus masas son comparables con la energía del experimento. A mayor distancia, la simetría entre el fotón y las partículas W y Z queda totalmente velada; en efecto, no hay energía suficiente para crear W o Z reales y, por tanto, no pueden observarse de una manera directa. Sólo pueden contemplarse los escasos efectos de corto alcance de su intercambio virtual (tales como la desintegración beta del neutrón). Este es el dominio actual de la física de partículas.

El concepto de rotura espontánea de simetría resuelve el problema de la conservación de la carga débil. A energías muy por encima de los 100 GeV, donde la simetría $SU(2) \times U(1)$ se observa directamente, resulta despreciable la masa de un quark o de un leptón; la helicidad de las partículas queda esencialmente fijada y, por tanto, la carga débil se conserva. A bajas energías, donde la simetría se rompe espontáneamente, no se conserva la carga débil, sino que puede desaparecer en el vacío cuando una partícula dotada de masa cambia de helicidad.

Esta teoría de las interacciones débiles y electromagnéticas fue elaborada entre los años 1960 y 1970 por Sheldon Lee Glashow y Steven Weinberg, de la Universidad de Harvard, y por Abdus Salam, del Centro Internacional de Física Teórica de Trieste (Italia). Glashow fue el primero en deducir la forma $SU(2) \times U(1)$ de la teoría, pero no supo cómo incorporar las masas de las partículas W y Z. Weinberg y Salam, cada uno por su lado, encontraron posteriormente la forma $SU(2) \times U(1)$ y aplicaron la noción de rotura espontánea de simetría, formulando así una teoría coherente.

La teoría $SU(2) \times U(1)$ es sólo una unificación parcial, puesto que incluye todavía dos fuerzas distintas, cada una con su propio grupo de simetría y su propia constante de acoplamiento. El cociente de las constantes de acoplamiento es un parámetro libre, que debe ser elegido para ajustar los datos experimentales. Otra deficiencia de la teoría es que la carga eléctrica se halla sólo parcialmente cuantificada. La construcción de dobletes de partículas relacionados por las transformaciones $SU(2)$ exige que todas las diferencias entre cargas eléctricas sean enteras, de forma que cada partícula pueda cambiar su identidad emitiendo un W^+ o un W^-. La carga media de los dobletes, sin embargo, no está cuantificada. La carga media es la carga $U(1)$, definida en conjunción con la carga electromagnética de la electrodinámica cuántica; como en esta teoría, no hay ninguna razón fundamental para que las cargas se limiten a valores enteros. En efecto, en los dobletes constituidos por quarks el intervalo entre las cargas es un entero, pero las cargas son fraccionarias.

La empresa de construir una teoría unificada de las fuerzas débiles, fuertes y electromagnéticas no debería verse como un intento de eliminar el modelo $SU(3)$ de color o el modelo $SU(2) \times U(1)$. Las teorías de las distintas fuerzas operan de una manera excelente para caer en la ligereza de suprimirlas. Una teoría unificada puede proporcionar una superestructura en la que se puedan englobar las teorías $SU(3)$ y $SU(2) \times U(1)$. La superestructura tendría la forma de una simetría mayor en la que quarks y leptones se hallarían en íntima relación.

La búsqueda de esta simetría mayor debe empezar por la búsqueda de un grupo mayor, que incluya tanto $SU(3)$ como $SU(2) \times U(1)$, como estructuras componentes. Aunque muchos grupos detentan esta propiedad, hay un candidato con muchos triunfos en su mano. Nos referimos al $SU(5)$, el grupo de todas las transformaciones posibles de cinco objetos distintos, o de una matriz cinco por cinco. De entre los grupos simples, es el menor que puede acomodar las simetrías constituyentes $SU(3)$ y $SU(2) \times U(1)$. En mi opinión, nos hallamos ante el grupo total de simetría de la naturaleza. Con Glashow propuse en 1973 una teoría unificada basada en el $SU(5)$.

En la representación más simple del grupo $SU(5)$, los cinco objetos son las componentes dextrógiras del quark d, en cada uno de los colores rojo, verde y azul, la componente dextrógira del positrón y la componente dextrógira del antineutrino de tipo electrónico (que tiene sólo la componente dextrógira). A cada uno de los cinco tipos de partículas se le asigna un valor para cada una de las cuatro cargas independientes: carga eléctrica, carga débil y dos cargas de color, que tomaré como R − V y V − A.

Veinticuatro partículas intermedias dan cuenta de todas las transiciones posibles entre estos cinco estados de la materia. Cuatro de las partículas son el fotón, el Z^0 y los gluones G_1 y G_2, que están directamente asociados con las cuatro cargas fundamentales. Al no llevar cargas, estas partículas sólo pueden tomar parte en aquellas interacciones en las que no cambia la identidad de las partículas. De las 20 restantes partículas intermedias, ocho nos son familiares: las partículas W^+ y W^-, que pueden convertir un positrón en un antineutrino y viceversa, y los seis gluones que transforman los colores de los quarks. Con este conjunto de 12 partículas intermediarias pueden explicarse todas las interacciones observadas hasta ahora en la naturaleza. El grupo $SU(5)$ incluye además otras 12 partículas intermediarias que son necesarias si pretendemos que la teoría goce de la máxima simetría posible. Las 12 partículas extra se designan por X, y son las mediadoras de la interconversión de leptones y quarks. Cada partícula X lleva carga débil, carga de color y carga eléctrica; las cargas eléctricas tienen valores de más o menos 1/3 y de más o menos 4/3.

Como ocurre con la distribución de las cargas de color en el $SU(3)$, la tabla que da las cargas en la teoría $SU(5)$ revela algunas irregularidades intrigantes. Para cada clase de carga, la suma de las cargas asignadas a las cinco partículas es nula. Por ejemplo, cada uno de los tres quarks coloreados tiene una carga eléctrica de $-1/3$, pero estas cargas están compensadas por la carga eléctrica del positrón, que es $+1$. Hemos de advertir que las cuatro variedades de carga son transportadas por al menos alguna de las partículas intermediarias del $SU(5)$. Los gluones tienen color, los W^+ y W^- poseen tanto carga débil como eléctrica y las partículas X llevan las cuatro formas de carga.

De estos dos hechos se puede deducir que todas las cargas están necesariamente cuantificadas. Las cargas eléctricas deben ser múltiplos de 1/3; si se aceptara en la familia una partícula con alguna otra carga, las partículas portadoras del $SU(5)$ no podrían ser emitidas o absorbidas por ella sin violar la conservación de la carga. Más aún, no es sólo el intervalo mínimo entre las cargas lo que queda fijado; los mismos valores de las cargas quedan determinados por la exigencia de que la carga total sea cero. Aquí hay, finalmente, una explicación de la cuantificación de la carga eléctrica. La misma exigencia explica la conmensurabilidad exacta de las cargas de los leptones y de los quarks, lo cual, a su vez, implica la neutralidad exacta del átomo. Además, de

la estructura de la familia se deduce una consecuencia intrigante: los sistemas con carga de color neutro tienen carga eléctrica entera.

¿Qué ocurre con las partículas restantes de la primera generación? Una de las características más elegantes de la teoría $SU(5)$ es que las cinco partículas dextrógiras de la familia menor de $SU(5)$ pueden combinarse en pares a fin de dar una familia de 10 partículas levógiras. Estos 10 estados componen la siguiente representación, en orden de simplicidad, del grupo. Son las componentes levógiras del quark d, el quark u y un antiquark $\bar{u}$ (en tres colores cada uno) y del positrón. Como en la construcción de los antiquarks y de los gluones, a partir del triplete básico de los quarks coloreados, este proceso no debe interpretarse como una prescripción física para construir las partículas. Un leptón o un quark levógiro no están, en realidad, compuestos como el estado ligado de dos partículas dextrógiras. Sin embargo, las 10 maneras posibles de formar pares de los cinco estados dextrógiros aportan todas las cargas correctas para las partículas levógiras.

Las transiciones posibles para cualquier partícula vienen también recogidas correctamente por este esquema de composición por pares. Incluyen tanto transiciones quark-leptón como quark-antiquark. Lo que aún es más importante, las transiciones no observadas no son permitidas por la estructura del grupo. Cada familia de partículas está cerrada, es decir, completa en sí misma. Cada transición permitida da origen a otra partícula de la misma familia y no hay más transiciones posibles.

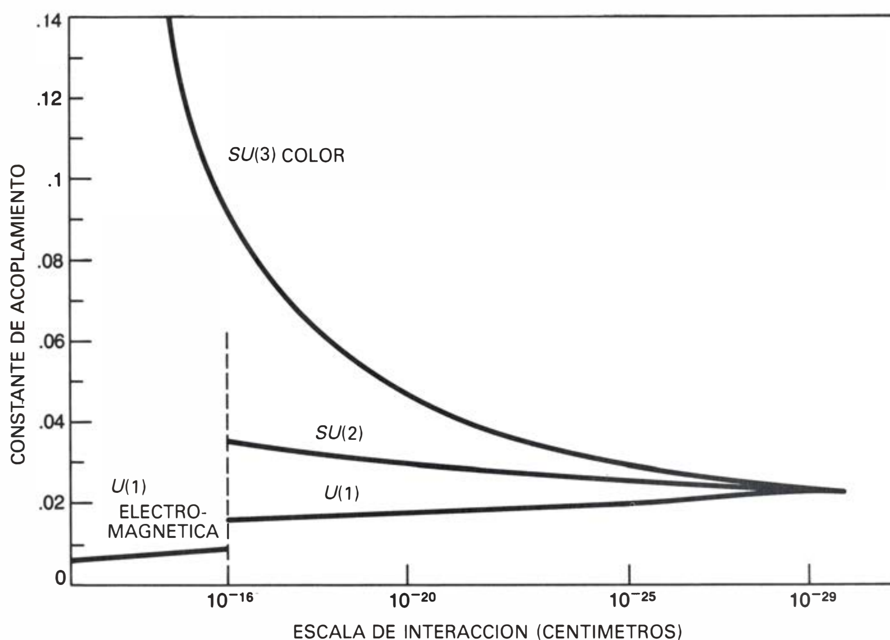
La familia de los cinco estados dextrógiros y la familia de los 10 estados levógiros suman un total de 15 partículas. Se pueden construir dos familias adicionales de forma ligeramente distinta para acomodar las restantes 10 partículas dextrógiras y 5 levógiras, que son las antipartículas de los 15 estados en las primeras dos familias. Por tanto, los 30 estados elementales de la primera generación tienen un lugar en la teoría y no hay lugares vacíos. Se pueden construir representaciones equivalentes de generaciones más altas sustituyendo el electrón por el muon o el leptón tau, el quark s o b por el quark d, etcétera.

Revisaremos lo hecho hasta ahora. En primer lugar, se tomó el grupo $SU(5)$ como grupo más pequeño, donde se podía englobar $SU(3)$ y $SU(2) \times U(1)$. A continuación, se eligieron cinco componentes dextrógiros de las partículas como miembros de la familia más simple de $SU(5)$. Las restantes componentes debían poder ajustarse en alguna otra familia de $SU(5)$, y así sucedió. Sin que sobre ni falte nada, quedan ajustadas en la familia siguiente en orden de simplicidad. Más aún, la composición de la familia derivada quedó especificada por un sencillo procedimiento de combinar las partículas en pares. Importa destacar que este procedimiento de combinación no tiene por qué funcionar. En muchos grupos distintos de $SU(5)$ no valdría. Representa el primero, el más sencillo y, en algunos aspectos, el éxito más notable de la teoría $SU(5)$.

El significado más obvio de la unificación $SU(5)$ estriba en que ya no son irreconciliablemente distintos los leptones y los quarks. Por el contrario, se trata de miembros de una única familia; un quark puede convertirse en un leptón (o viceversa) con idéntica facilidad con que un quark puede convertirse en un quark distinto o un leptón en otro leptón distinto. Una consecuencia más de la unificación es que las fuerzas débiles, fuertes y electromagnéticas tendrían todas la misma intensidad, o la misma constante de acoplamiento. Ninguna de estas predicciones se satisfacen en el mundo tal como aparece hoy día. En los millones de interacciones de partículas elementales registradas por los físicos no se sabe de ningún caso en que se haya observado una conversión leptón-quark. Más aún, las constantes de acoplamiento de las tres fuerzas difieren de un modo notable: la fuerza fuerte es aproximadamente cien veces más intensa que el electromagnetismo, hallándose la fuerza débil entre ambas. Entonces, si $SU(5)$ es una simetría de la naturaleza, es evidente que está muy rota.

La simetría podría romperse mediante un mecanismo similar al que destruye la simetría $SU(2) \times U(1)$ de las fuerzas débiles y electromagnéticas. De esta forma, las partículas X adquirirían una gran masa y los efectos debidos al intercambio de las partículas X quedarían suprimidos. Sin embargo, en el $SU(5)$ la rotura debería presentarse a energías mucho más altas o, lo que es equivalente, a distancias mucho más pequeñas que en $SU(2) \times U(1)$. Esta distancia es la escala de unificación, la distancia a la cual se manifiesta toda la simetría de la teoría.

Con la teoría $SU(5)$ se puede especular cómo aparecería el mundo a distintas escalas de distancias o energías. En



EN LA TEORIA $SU(5)$, se espera que la unificación de fuerzas se patentice a energías extraordinariamente altas o, de manera equivalente, a una distancia muy pequeña. La intensidad intrínseca de cada fuerza viene medida por una constante de acoplamiento sin dimensiones, asociada con la simetría subyacente. A causa de los efectos de las partículas virtuales que rodean una carga, la interacción fuerte $SU(3)$ disminuye con la distancia cuando se acorta la separación de las partículas. La interacción $SU(2)$ también se hace más débil a distancias pequeñas, aunque ello ocurre más lentamente. La interacción $U(1)$, por el contrario, aumenta su intensidad. Una extrapolación de los valores de las constantes medidas a distancias comparativamente altas sugiere que las curvas convergen a una distancia de unos 10^{-29} centímetros, equivalente a una energía de 10^{15} gigaelectronvolt. A esta distancia y energía, todas las fuerzas tendrían la misma intensidad y serían igualmente probables todas las interacciones entre las partículas elementales. A distancias mayores de 10^{-16} centímetros se rompe la simetría $SU(2) \times U(1)$ y las fuerzas $SU(2)$ y $U(1)$ deja de existir como entidades separadas. La simetría $U(1)$ del electromagnetismo permanece, pero la fuerza débil aparece solamente a través del intercambio virtual entre la partícula W y la partícula Z.

un experimento que estudiara distancias mucho menores que la escala de unificación, la invariancia gauge $SU(5)$ del mundo se manifestaría claramente. Todas las interacciones, incluidas las propias transformaciones leptón-quark y quark-antiquark, estarían en pie de igualdad, pues todas las partículas intermediarias del $SU(5)$ (el fotón, los gluones, las partículas W y Z y las partículas X) se crearían, esencialmente, con la misma probabilidad. Las masas de las partículas W, Z y X apenas si servirían para distinguirlas del fotón y de los gluones, toda vez que serían pequeñas en comparación con la energía del experimento.

A distancias parecidas a la escala de unificación se observaría la compleja física asociada con la rotura espontánea de la simetría $SU(5)$. Las partículas X serían emitidas, pero su masa las haría muy distintas de todas las demás partículas. A distancias muy por encima de la escala de unificación (aunque lejos todavía de alcanzar los 10^{-16} centímetros) la simetría $SU(5)$ estaría casi totalmente oculta. Las partículas X no podrían crearse como partículas reales; y, por tanto, leptones y quarks se distanciarían en familias diferentes, con poca comunicación entre sí. Por otro lado, la simetría $SU(2) \times U(1)$ permanecería intacta, de manera que se podrían hacer pocas distinciones entre las interacciones débiles y electromagnéticas. A distancias mayores que 10^{-16} centímetros, se truncaría también la simetría $SU(2) \times U(1)$, y habría tres fuerzas distintas.

La rotura espontánea de la simetría puede explicar también las disparidades entre las constantes de acoplamiento. El elemento crucial de la explicación reside en el efecto de las partículas virtuales en el vacío que rodea a una carga puntual. En el caso de la fuerza fuerte, como se recordará, la nube de gluones virtuales que envuelve un quark esparce de forma efectiva la carga de color, por cuyo motivo la constante de acoplamiento decrece cuando se mide a distancias más pequeñas. Las partículas W virtuales ejercen un efecto similar sobre la carga débil, aunque algo menor al haber menos partículas W que llevan carga débil que gluones que porten carga de color. En la teoría $U(1)$, por otra parte, el hecho de que el Z^0 no tenga carga origina un fenómeno muy distinto. A causa de la polarización de los electrones y positrones virtuales, la constante de acoplamiento $U(1)$ aumenta a distancias menores.

De estas tendencias de las constantes de acoplamiento podemos deducir una consecuencia sencilla. A grandes distancias, la constante de acoplamiento del $SU(3)$ de la fuerza fuerte es la mayor, pero también la que decrece más deprisa; la constante $SU(2)$ de las fuerzas débiles es más pequeña y decrece más lentamente; la constante de $U(1)$ es la menor, si bien aumenta al disminuir la distancia. Por tanto, parece que puede existir una distancia en la que las tres constantes de acoplamiento tengan, aproximadamente, el mismo valor.

A distancias menores de 10^{-16} centímetros, todas las constantes de acoplamiento son bastante pequeñas y puede calcularse la forma como varían con la distancia o la energía. Se pueden determinar los valores a distancias progresivamente menores; la distancia a la que las tres constantes convergen es la escala de unificación. Si se estudian así dos de estas constantes, el valor de la tercera puede predecirse a cualquier energía. Tales cálculos se han hecho usando como datos la constante de acoplamiento fuerte y la constante de acoplamiento electromagnética; la última vuelve a ser una combinación de las constantes de acoplamiento de $U(1)$ y $SU(2)$. Los resultados suministran valores para la escala de unificación y para el cociente de las constantes de $U(1)$ y $SU(2)$, un parámetro arbitrario en las teorías no unificadas.

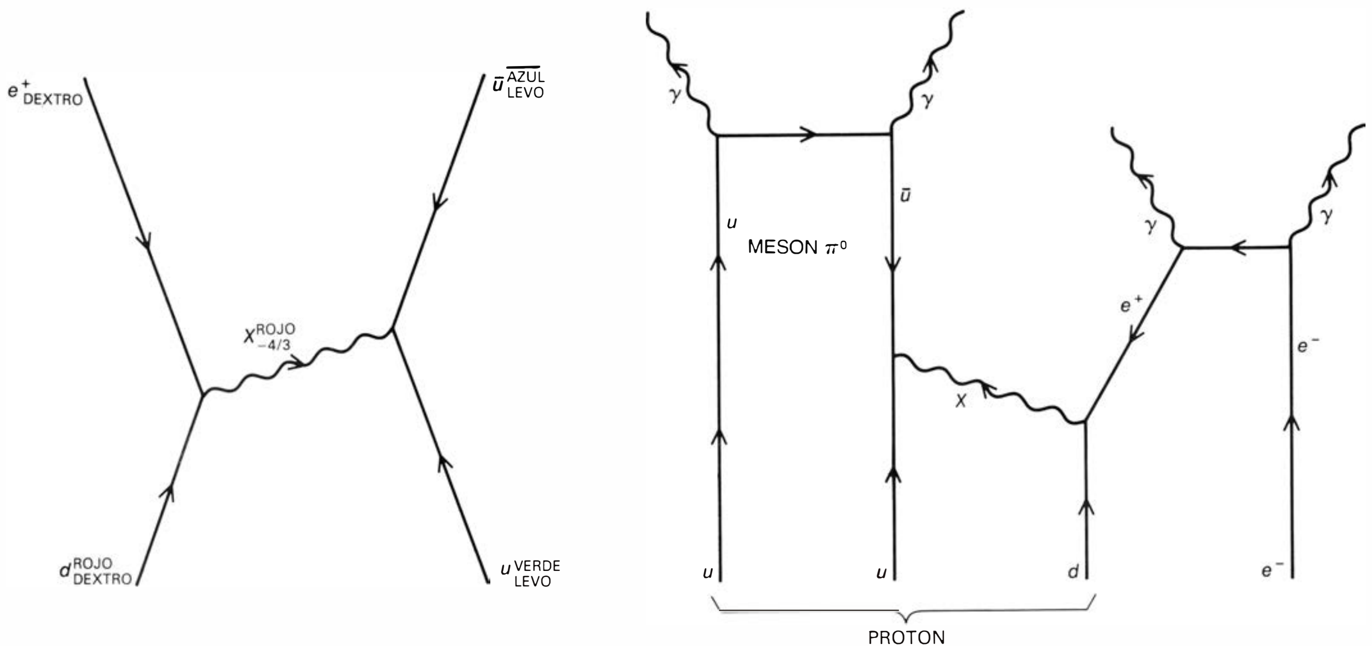
En 1974, hice este cálculo con Helen R. Quinn (que está ahora en el Centro del Acelerador Lineal de Stanford) y con Weinberg, para una clase de teorías unificadas que incluye el $SU(5)$. Obtuvimos una distancia de unos 10^{-29} centímetros para la escala de unificación y un valor de 0,2 para el cociente de la constante de $U(1)$ y la de $SU(2)$. Por aquel entonces, los resultados no eran alentadores, ya que las medidas del cociente sugerían un valor en torno a 0,35. Posteriores mediciones más precisas del cociente han dado valores más bajos. Hoy se sostiene que el cociente vale 0,22 más o menos 0,02, en acuerdo con el resultado teórico.

Otra predicción del modelo $SU(5)$ comprobable a las energías accesibles fue estudiada, en 1977, por Andrzej Buras, John Ellis, Mary K. Gaillard y Demetres V. Nanopoulos, de la Organización Europea de Investigaciones Nucleares (CERN), de Ginebra. Se encontraron con que, en la versión más simple de $SU(5)$, se podía determinar el cociente de la masa del quark b y de la masa del leptón tau. Como en el caso de las constantes de acoplamiento, se espera que las masas sean iguales a la escala de unificación; a distancias mayores, sin embargo, el quark b es más pesado en razón de su carga de color. Se calculó el cociente de masas a baja energía, cifrándose en torno a 3:1. La masa del tau es conocida: casi dos veces la masa del protón. No se tiene la misma certeza respecto a la masa del quark b, ya que el quark no puede examinarse aisladamente. La mejor estimación de que disponemos hoy la sitúa en torno a cinco veces la masa del protón, por lo que el cociente de masas será de 5:2.

La escala de unificación de 10^{-29} centímetros es una distancia extraordinariamente pequeña. (Si un protón se hinchara hasta alcanzar el tamaño del sol no alcanzaría aún el micrometro.) Implícita en la unificación $SU(5)$ está la hipótesis de que no aparecen nuevos principios físicos en todo el intervalo de distancias entre 10^{-16} y 10^{-29} centímetros; en particular, se debe suponer que la forma en que las constantes de acoplamiento varían con la distancia no cambia. Tal hipótesis es, por supuesto, molesta, pero no totalmente implausible. Hay ya una escala pequeña de distancias a la que se esperan nuevos fenómenos. A unos 10^{-33} centímetros la gravitación puede llegar a ser tan fuerte como las otras fuerzas y, por tanto, cualquier teoría que describa sucesos a alta escala habrá de incluir la gravitación. Me parece alentador que la escala de unificación, aunque pequeña, sea 10.000 veces mayor que 10^{-33} centímetros.

Una distancia de 10^{-29} centímetros corresponde a una energía de alrededor de 10^{15} GeV o, aproximadamente, 10^{15} veces la masa del protón. Las partículas X deben tener una masa equivalente a esta energía. A modo de comparación, diremos que las partículas más pesadas que pueden hoy día crearse con aceleradores de partículas tienen una masa de alrededor de 10 GeV. Se confía en que la nueva generación de aceleradores en proyecto permita alcanzar los 100 GeV que se necesitan, aproximadamente, para crear las partículas W y Z. Para crear las partículas X la energía debería incrementarse en 13 órdenes de magnitud, lo que parece poco probable que pueda conseguirse algún día.

Aun cuando nunca fuera posible exhibir una partícula X real en el laboratorio, la existencia de las mismas puede demostrarse detectando sucesos en los que se intercambia una partícula X virtual. Tales intercambios serían también



DESINTEGRACION DEL PROTON: consecuencia inevitable de las transiciones *SU*(5), que convierten un quark en un leptón o un quark en un antiquark (*izquierda*). En primer lugar, un quark *d* rojo y dextrógiro emite una partícula *X* roja con carga eléctrica −4/3; en virtud de lo cual, el quark se transforma en un positrón dextrógiro. A continuación, la partícula *X* es absorbida por un quark *u* verde y levógiro, que se convierte en un antiquark $\bar{u}$ antiazul y levógiro. El diagrama de la derecha muestra el resultado del proceso cuando ocurre en un protón, que forma el núcleo de un átomo de hidrógeno. El protón tiene una composición *uud*. Como resultado del intercambio de la partícula *X*, el quark *d* se convierte en un positrón, que puede acabar por colisionar con el electrón del átomo (o con algún otro electrón), aniquilándose ambas partículas en un impulso de fotones de alta energía. Lo que queda del protón después de intercambiarse la partícula *X* es un quark *u* y un antiquark $\bar{u}$ que, juntos, constituyen un mesón pi neutro. Este se desintegra también en fotones de alta energía. El resultado final del intercambio es la conversión de toda la masa del átomo de hidrógeno en radiación electromagnética.

muy raros, ya que podrían sólo presentarse cuando las dos partículas elementales se acercasen a distancias menores de 10^{-29} centímetros, la una de la otra. Incluso en la confusión de sucesos más usuales, el intercambio de una partícula *X* se advertiría fácilmente, pues las partículas *X* pueden hacer algo vedado a las otras: transformar un quark en un leptón o un quark en un antiquark. Este proceso cuestiona la estabilidad misma de la materia.

Las interacciones mediadas por las partículas *X* se distinguen de las restantes interacciones en que violan la conservación de una cantidad llamada número bariónico. El número bariónico de cualquier partícula puede definirse como un tercio del número de quarks menos un tercio del número de antiquarks. De aquí que el protón y todas las restantes partículas formadas por tres quarks tengan número bariónico +1, mientras que el mesón pi y las otras partículas que constan de un quark y un antiquark poseen número bariónico 0. Por supuesto, los leptones tienen también número bariónico 0, ya que no están formados ni por quarks ni por antiquarks. En las interacciones fuertes, débiles y electromagnéticas, el número bariónico total no puede cambiar nunca. Si la conservación del número bariónico fuera una ley absoluta de la naturaleza, el protón no podría desintegrarse nunca, dado que el protón es la partícula más ligera con número bariónico no nulo. La unificación *SU*(5) predice, sin embargo, que el protón se desintegra.

Un fallo posible de la conservación del número bariónico puede ilustrarse en un protón que forma el núcleo de un átomo de hidrógeno. El protón consta de los quarks *u, u* y *d,* con un quark en cada uno de los tres colores, Si dos de los quarks se aproximan a una distancia de 10^{-29} centímetros, una partícula *X* puede pasar de uno al otro. Por ejemplo, un quark *d* rojo dextrógiro puede emitir un *X* con carga eléctrica −4/3 y cargas de color correspondientes al color rojo. El quark *d,* al haber perdido su carga de color, y haber cambiado su carga eléctrica de −1/3 a +1, se convierte, por tanto, en un positrón. Por otra parte, la partícula *X* puede ser absorbida por un quark *u* verde levógiro, que se convertiría en un antiquark $\bar{u}$ levógiro, con el color antiazul. El nuevo antiquark $\bar{u}$ se combinaría con el restante quark *u* para formar un mesón pi neutro. Los números bariónicos, tanto del positrón como del mesón pi, son nulos, de forma que el número bariónico total cambia de +1 a 0.

Si se observase este proceso en el laboratorio, el intercambio de la partícula *X* no se podría percibir directamente. Todo lo que se vería sería la desintegración de un protón en un positrón y un mesón pi neutro. Pero no concluiría aquí la secuencia de procesos. El positrón encontraría, posteriormente, un electrón (quizás el electrón del átomo de hidrógeno) y se aniquilarían entre sí, dando origen a rayos gamma, es decir, a fotones de alta energía. El quark *u* y el antiquark $\bar{u}$ del mesón neutro pi acabarían aniquilándose entre sí de forma análoga, dando origen a nuevos rayos gamma. El resultado último sería que un átomo de hidrógeno se desintegraría en un estado de radiación pura. Este proceso representa una conversión de materia en energía mucho más eficiente que la fisión nuclear o fusión termonuclear. La fusión de los átomos de hidrógeno para formar helio libera menos del 1 % de su masa como energía, mientras que el proceso que acabamos de describir libera el 100 por ciento de la masa.

La desaparición brusca de un protón y, por tanto, de un átomo es un acontecimiento que debe suceder muy raramente; si así no fuera, se habría descubierto hace años. En efecto, se supone que habrá una frecuencia muy baja, ya que las partículas casi nunca se sitúan a una distancia donde sea posible el in-

tercambio de una partícula *X*. Quinn, Weinberg y el autor emplearon su cálculo de la escala de unificación para estimar el ritmo de desintegración del protón. Nuestra estimación ha sido revisada por muchos otros, de los que puedo citar a Buras, Ellis, Gaillard y Nanopoulos, del CERN, Terrence J. Goldman y Douglas A. Ross, del Cal Tech, y William J. Marciano, de la Universidad Rockefeller. La estimación actual cifra la vida media del protón en unos 10^{31} años.

Evidentemente, resulta inviable esperar 10^{31} años para que un determinado protón se desintegre, y así confirmar la validez de la unificación *SU*(5). Adviértase que la edad del universo desde la explosión inicial es de sólo uno 10^{10} años. Pero sí cabe la búsqueda de la desintegración del protón. Una vida media de 10^{31} años implica que en una colección de 10^{31} protones se debe observar una desintegración cada año. En 1000 toneladas de materia hay unos 5×10^{32} protones y neutrones, de manera que se puede esperar que 50 de ellos se desintegren cada año. De aquí que la estrategia para detectar sucesos que violen la conservación del número bariónico consista en controlar todo lo que pasa en, al menos, 1000 toneladas de materia durante varios años y distinguir las desintegraciones de protones y neutrones de sucesos más comunes.

Hay varios grupos de investigación empeñados en la proyección de experimentos a esta escala. Se harán éstos a gran profundidad, en la tierra o en el agua, para reducir al mínimo el número de rayos cósmicos que atraviesen la muestra de materia considerada. Los rayos cósmicos quizás originaran interacciones que podrían confundirse con la desintegración de un protón. Se montará un experimento en una mina de sal cerca de Cleveland; otro, en una mina de plata en Utah y un tercero, en una mina de hierro en Minnesota. Se han proyectado también experimentos a una escala ligeramente menor para dos túneles debajo de los Alpes y experimentos más pequeños se están llevando a cabo ya en minas de oro de Dakota del Sur y la India.

La energía necesaria para crear partículas *X* reales tal vez trascienda siempre las capacidades de las máquinas hechas por los hombres y cabe incluso la posibilidad de que en el universo actual no exista ningún proceso por el que generar una energía tan alta. Pero nada hay contra la posibilidad de que, en épocas precedentes, abundaran las partículas *X*. Unos 10^{-40} segundos después de la explosión inicial el tamaño del universo era comparable a la escala de unificación. La temperatura del universo era entonces tan elevada (alrededor de 10^{18} grados Kelvin) que todas las partículas tenían energías comparables a la masa de la *X*. Por consiguiente, la simetría *SU*(5) empezaba a romperse y las conversiones quark-leptón eran tan frecuentes como cualquier otra interacción. No había diferencia fundamental entre los quarks y los leptones, o entre las fuerzas fuertes, débiles o electromagnéticas: había una sola clase de materia y una sola fuerza.

Un resto de aquel período, de manifiesta simetría, puede existir en el universo actual; en cierto sentido, el universo actual es el resto. Una antigua paradoja que la astrofísica tiene planteada es la de por qué el universo está constituido de materia más que de antimateria. Parece que sería de esperar que hubiera cantidades iguales de materia y de antimateria, que acabarían aniquilándose entre sí, quedando al final un universo que sólo tendría radiación. La unificación *SU*(5) ofrece una explicación coyuntural para el aparente predominio de la materia. Es posible que el libre intercambio de partículas *X* en un breve período después de que la simetría *SU*(5) se rompiera creara más quarks que antiquarks y, por tanto, más bariones que antibariones.

La intrigante especulación de que los procesos que violan el número bariónico pueden ser los responsables del exceso de bariones fue desarrollada, por primera vez, por el físico ruso Andre Sakharov, en 1967. Más recientemente, Motohiko Yoshimura, de la Universidad de Tohoku, sugirió que la violación del número bariónico predicho por las teorías unificadas tenía las propiedades correctas para explicar el exceso observado. La idea ha sido elaborada entre otros, por Ellis, Gaillard, Nanopoulos, Steven Weinberg, Savas Dimopoulos y Leonard Susskird, de la Universidad de Stanford, y Sam B. Treiman, Anthony Zee y Wilczek, de Princeton. Han demostrado que un exceso de bariones sobre antibariones puede originarse sólo si los procesos que violan la conservación del número bariónico aparecen de forma distinta cuando se invierte el sentido del transcurso del tiempo. Esta condición se satisface en la teoría *SU*(5). Así pues, un ápice de evidencia de la unificación *SU*(5), aunque sea una indicación indirecta y circunstancial, es la misma existencia de materia.

Reconocimiento del habla por medio de ordenadores

Diseñar una máquina que escuche es mucho más difícil que construir una que hable. Sólo con una mejor comprensión de los modelos humanos del habla habrá progresos significativos en su reconocimiento automático

Stephen E. Levinson y Mark Y. Liberman

Los modernos ordenadores tienen poderes prodigiosos, pero resultarían aún más útiles si pudiéramos disponer de medios más naturales para comunicarnos con ellos. La evolución ha hecho el lenguaje hablado muy acorde con las necesidades de la comunicación humana. Es rápido y casi no requiere ningún esfuerzo. Tampoco necesita contacto visual o físico y apenas limita la movilidad del cuerpo o el uso de las manos. Una máquina capaz de reconocer el habla humana podría combinar todas estas ventajas con los poderes bien diferentes del ordenador. Una máquina así podría proporcionar un acceso universal a las grandes bases de datos a través de la red telefónica. Podría tomar el control de máquinas complejas mediante órdenes orales y hacer posibles refinados artilugios protéticos para minusválidos.

Sin embargo, aun después de más de 40 años de investigación, el reconocimiento automático del habla natural o conversacional sigue siendo un objetivo utópico. Los sistemas actuales para el reconocimiento del habla comprenden un vocabulario corto y disponen de poca capacidad para manejar secuencias fluidas de palabras: por lo general, hay que adiestrarlos para reconocer tan sólo la voz de un único hablante. Y aun así, las ventajas del reconocimiento automático del habla son tan grandes, que ya se han comercializado aparatos, económicamente prácticos en determinadas aplicaciones, capaces de reconocer palabras aisladas o frases cortas a partir de un vocabulario que va de 10 a 30 entradas. En los laboratorios de investigación hay reconocedores de habla con vocabularios que alcanzan hasta 1000 palabras, sistemas que reconocen oraciones a partir de un vocabulario limitado con pausas breves entre las palabras e incluso sistemas que reconocen el habla conexa con bastante precisión si el vocabulario es corto, la sintaxis reducida y el hablante cuidadoso.

La interacción de la tecnología y la economía llevará indudablemente a sistemas de reconocimiento del habla con mayor capacidad. No podemos predecir con exactitud la marcha de este desarrollo, pero tenemos, en cambio, la seguridad de que la mera elaboración y las estimaciones de la tecnología actual no conducirán al desarrollo de máquinas equiparables a la facultad humana de reconocer el habla. El principal progreso depende de nuevos descubrimientos.

¿A qué se debe que el problema del reconocimiento sea tan peliagudo? El meollo de la dificultad reside en los complejos y variables medios con que los mensajes lingüísticos se codifican en el habla. La lengua hablada permite expresar los pensamientos en forma de sonidos y captar mensajes a partir de los sonidos que otros emiten. Este curioso dispositivo bidireccional entre conceptos mentales y vibraciones aéreas presupone que los interlocutores tengan en común un cierto entramado conceptual, a fin de que el mensaje recibido sea, al menos aproximadamente, equivalente al emitido. Pero no basta compartir el conocimiento de las cosas que uno quisiera decir. Seguramente los hablantes monolingües de inglés y de finés tienen muchos mensajes potenciales en común y sin embargo no pueden comprender las respectivas enunciaciones de sus oponentes. Hablar y comprender requiere, además, compartir un sistema común que codifique mensajes en sonidos y descodifique sonidos de habla para dar lugar a significados. En otras palabras, hay que conocer la misma lengua.

La comunicación hablada por medio de ordenadores presenta características análogas. El ordenador "sabe" (en un sentido un tanto ampliado de la palabra) lo mismo que sus usuarios acerca de un determinado dominio. Conviene, entonces, intercambiar información en este dominio y resulta que el habla es el medio de comunicación elegido.

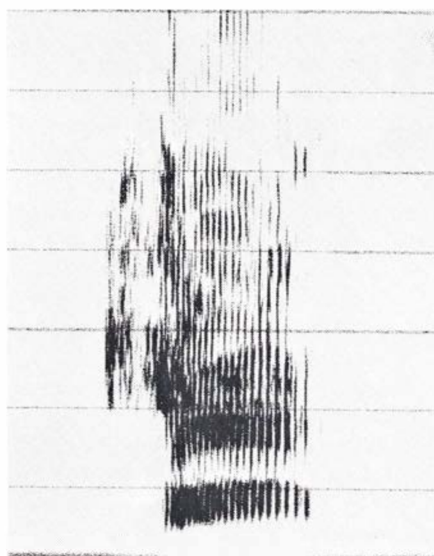
Considérese una conversación entre un ordenador y sus operadores sobre el inventario de existencias en un almacén. El ordenador "sabe" qué cantidad de un determinado género está disponible y dónde está depositado cada artículo. Su base de datos consigna asimismo costos y proveedores. Es probable que la gente conciba de muchas maneras el almacén y su contenido, pero la estructura de la base de datos del ordenador es similar al pensamiento humano, al menos para posibilitar ciertos tipos de comunicación. Los operadores plantean preguntas al ordenador, en principio asequibles, tales como "¿Disponemos de lápices azules?". Hay también cosas que el ordenador puede "comprender" provechosamente, como "Ya no queda espacio en la nave 13". Si esta suerte de comunicación ha de realizarse por medio del habla, es menester que el ordenador y sus usuarios se pongan de acuerdo sobre el modo de codificar dichos mensajes en sonidos y de invertir el proceso. Han de "conocer", en consecuencia, la misma lengua.

Si nos ocupamos principalmente de lenguas como el inglés [o el español], llamadas naturales, es porque quedan implícitamente definidas por el uso diario de la gente corriente. Actualmente, los ordenadores funcionan con lenguajes formales como el *fortran*, compuestos mediante un conjunto explícito de reglas especialmente estipuladas por peritos. Al menos por ahora, los ordenadores sólo ejecutan aquello para lo que están programados. No viven en el mundo de los humanos ni aprenden nada a partir de la experiencia cotidiana.

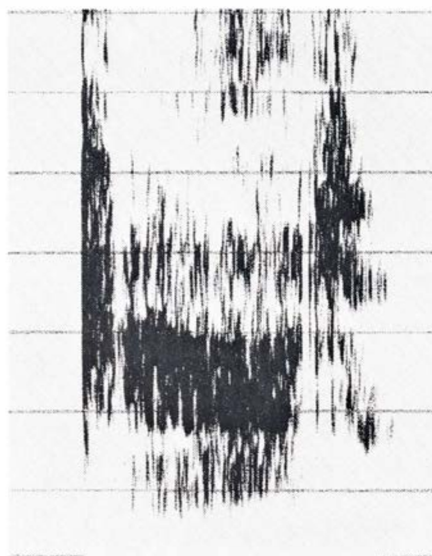
De ahí que, para que un ordenador “conozca” una lengua natural, hay que facilitarle una caracterización explícita y precisa de la misma, o al menos de lo que el programador entiende que es la lengua. En todos los sistemas existentes de reconocimiento del habla actualmente concebidos la descripción formal de una lengua natural sólo cubre un fragmento de la lengua en cuestión. A su vez, el formalismo reconstruye dicho fragmento sirviéndose de unos recursos probablemente muy distintos de los que utiliza el conocimiento implícito de un hablante nativo. Pero, aun siendo imperfectas las facultades lingüísticas del ordenador, bastan, no obstante, para posibilitar una provechosa comunicación con las personas.

A fin de comprender algunos intentos encaminados a lograr el reconocimiento de las lenguas naturales, conviene empezar por considerar algunos aspectos de la lengua y el habla en sus propios términos. Examinaremos luego ciertos métodos para reconocer palabras aisladas, así como algunos procedimientos para el análisis del habla conexa. Por último, describiremos un sistema de reconocimiento del habla, construido en los Laboratorios Bell, con el que se intenta combinar los principales elementos de la comunicación humana por medio del habla en una sola unidad operativa.

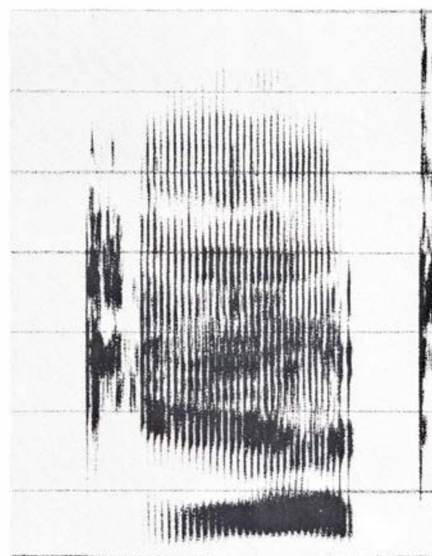
El núcleo del habla humana está constituido por la palabra. Las secuencias de palabras suelen disponerse en frases de acuerdo con unos principios combinatorios conocidos por el nombre de sintaxis. Se supone, además, que estas secuencias suelen construirse con el objeto de significar algo. El hecho de que las palabras formen parte normalmente de un discurso coherente puede tener su utilidad en el reconocimiento de las mismas, por cuanto proporcionan un contexto en el que ciertas palabras son más probables que otras. Ahora bien, es extremadamente difícil



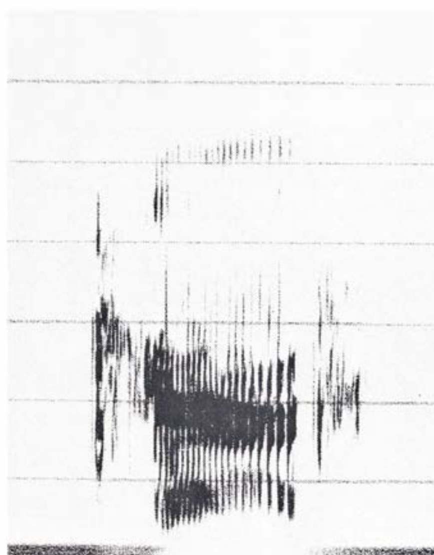
"CAT" [KæT]
("GATO") HABLANTE 1 MICROFONO



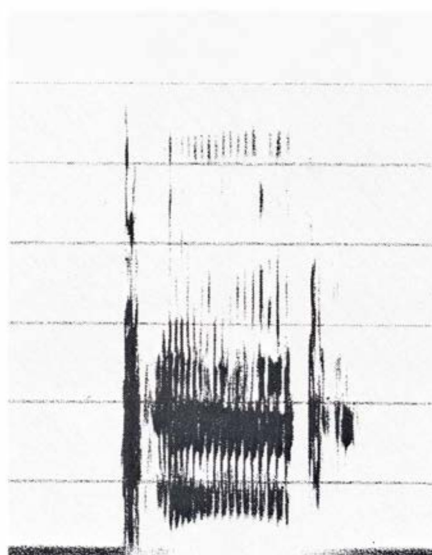
"CAT" [KæT]
("GATO") HABLANTE 1 CUCHICHEADO



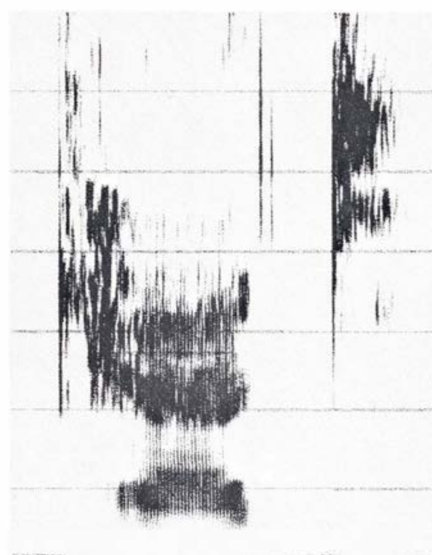
"CAT" [KæT]
("GATO") HABLANTE 2 MICROFONO



"CAT" [KæT]
("GATO") HABLANTE 1 TELEFONO



"PAT" [PæT]
("PALMADA") HABLANTE 1 TELEFONO



"CAT" [KæT]
("GATO") HABLANTE 3 MICROFONO

VERSATILIDAD del habla, ilustrada aquí por espectrogramas; constituye una de las principales dificultades con que tropieza la construcción de un sistema automático para el reconocimiento del habla. Los espectrogramas de palabras distintas pero acústicamente similares pueden resultar más parecidos que los espectrogramas de la misma palabra pronunciada en condiciones diversas por hablantes diferentes. El reconocimiento automático del habla debe ser capaz de atender tan sólo a las diferencias espectrales pertinentes (cuando existen) y soslayar aquellas que son lingüísticamente irrelevantes. Los espectrogramas de sonido representan una serie de espectros de amplitud a lo largo del tiempo. El factor tiempo discurre por el eje horizontal y la frecuencia por el vertical. Cuanto más oscura es la mancha del gráfico mayor es la amplitud de la onda en el momento y la frecuencia respectivos.

```
NFRAM = 396
NUM. DE PALABRAS = 5
CANDIDATOS PARA LA PALABRA NUM.  1  20 FRAMES
WHAT ("CUAL")        1     0,180
CANDIDATOS PARA LA PALABRA NUM.  2  29 FRAMES
IS ("ES")            1     0,270
CANDIDATOS PARA LA PALABRA NUM.  3  24 FRAMES
NINE ("NUEVE")       1     0,343
ONE ("UNO")          1     0,278
SIX ("SEIS")         1     0,370
SEVEN ("SIETE")      1     0,242
THE(ALT) ("LA")      1     0,314
CANDIDATOS PARA LA PALABRA NUM.  4  24 FRAMES
SOME ("ALGUN")       1     0,296
FLIGHT ("VUELO")     1     0,341
ONE ("UNO")          1     0,414
THREE ("TRES")       1     0,390
CANDIDATOS PARA LA PALABRA NUM.  5  28 FRAMES
NINE ("NUEVE")       1     0,378
OH ("OH")            1     0,390
TIME ("HORA")        1     0,291
ONE ("UNO")          1     0,411
SEVEN ("SIETE")      1     0,396
TIMES ("VECES")      1     0,315
WHAT IS SEVEN SOME TIME ("CUAL ES SIETE ALGUNA VEZ")
METRIC = 0,1279442E 01
WHAT IS THE FLIGHT TIME ("A QUE HORA ES EL VUELO")
METRIC = 0,1396801E 01
THE FLIGHT TIME IS FIVE HOURS AND TWENTY-FIVE MINUTES ("LA HORA DEL VUELO ES A LAS
CINCO HORAS Y VEINTICINCO MINUTOS")
```

Máquina: Aquí el sistema de información de vuelos de los Laboratorios Bell. ¿En qué puedo servirle?
Operador: Quiero un asiento de primera clase para el vuelo número treinta y uno a Denver, el domingo.
Máquina: Lo siento, pero este vuelo no se hace los domingos.
Operador: Quisiera salir el sábado.
Máquina: El vuelo número treinta y uno sale de Nueva York a las nueve cuarenta y cinco de la mañana y llega a Denver a la una y diez de la tarde.
Operador: ¿Cuánto dura el vuelo?
Máquina: El vuelo dura cinco horas y veinticinco minutos.
Operador: ¿Cuántas escalas hay en el vuelo?
Máquina: Este vuelo tiene una escala.
Operador: Quiero un vuelo directo.
Máquina: Lo siento, pero no hay vuelos directos a las nueve de la mañana.
Operador: Quiero ir a las cuatro de la tarde.
Máquina: El vuelo número dieciséis sale de Nueva York a las cuatro cincuenta de la tarde y llega a Denver a las siete de la tarde.
Operador: ¿Cuántas escalas hay en este vuelo?
Máquina: Este vuelo no tiene ninguna escala.
Operador: ¿Se sirve alguna comida durante el vuelo?
Máquina: Se sirve la cena.
Operador: ¿Cuánto es la tarifa?
Máquina: La tarifa en primera clase ida y vuelta es de trescientos sesenta dólares.
Operador: ¿Qué aparato hace el vuelo?
Máquina: El avión es un Douglas DC diez.
Operador: Prefiero el Boeing siete cuatro siete.
Máquina: El avión es un Douglas DC diez.
Operador: Por favor, repítame la hora de salida.
Máquina: La hora de salida es a las cuatro cincuenta de la tarde.
Operador: Tomaré el vuelo número sesenta y cinco.
Máquina: Lo siento, pero no hay ningún vuelo sesenta y cinco a Denver.
Operador: Por favor, repítame el número del vuelo.
Máquina: El número del vuelo es el dieciséis.
Operador: Quiero el vuelo dieciséis.
Máquina: Confirmada su reserva en el vuelo número dieciséis a Denver para el sábado. Gracias.

CONVERSACION telefónica entre el hombre y la máquina. Discurre siguiendo unos pasos que pueden representarse en la pantalla de un terminal de ordenador. El ordenador cuenta el número de palabras que hay que reconocer en cada oración y divide también el enunciado entero en "frames", o "armazones", tomando una muestra de espectro cada 15 milisegundos. Aquí, "*nfram*" es el número de armazones de la oración. Si excede el número de armazones ocupado por palabras concretas es porque el hablante ha de hacer una breve pausa entre las palabras. Las presuntas palabras enumeradas para cada posición oracional se han encontrado por comparación con las plantillas de palabras almacenadas en el ordenador. Sólo aparecen aquellas presuntas palabras gramaticalmente posibles en una posición dada y similares en estructura espectral a la emitida. No se consignan todas las posibles palabras que cabe considerar. Los números que siguen a las palabras candidatas representan la distancia que hay entre la plantilla de la palabra y el enunciado. Cuanto más corta es la distancia, más parecida es la plantilla al enunciado. La expresión "*metric*" se refiere a la suma no redondeada de las medidas de distancia para una sarta de palabras. Si el *metric* más pequeño posible (que consta necesariamente de la palabra más probable en cada posición) no está permitido por la gramática interna, queda sustituido por el siguiente *metric* más pequeño, gramaticalmente correcto. Al fin se da por teléfono, con voz sintetizada, una respuesta a la pregunta formulada por el operador. La conversación queda transcrita por el impresor del ordenador.

hacer que un ordenador actúe como si fuese capaz de "comprender" las secuencias de palabras. El problema no sólo afecta a las relaciones entre las palabras, sino también a la tarea de conocer y razonar acerca de la naturaleza del mundo.

Aunque la capacidad de comprender la lengua puede ser el fin último, la empresa del reconocimiento del habla se funda realmente en la identificación de las palabras. Y la parte de las palabras que nos interesa aquí es su sonido. A este respecto, una palabra es una clase de equivalencia de ruidos: el conjunto de todos los sonidos que aun siendo diferentes desde ciertos puntos de vista representan (en el contexto de su enunciación) una misma unidad léxica. El problema del reconocimiento de la palabra consiste en hallar un espacio matemáticamente definido en el que pueda delimitarse efectivamente tal conjunto de sonidos. Puesto que el grado de variación dentro del conjunto de sonidos correspondientes a una palabra dada es muy amplio, muy pequeña la distinción acústica entre palabras diferentes y como, en fin, el hablante adulto normal puede disponer de unas 100.000 palabras o más, el problema resulta verdaderamente peliagudo.

Para comprender las fuentes de variación en el sonido de una palabra y la naturaleza de la distinción entre una palabra y otra, es necesario asimilar un par de cosas. Primero, hay que comprender el medio básico de la comunicación hablada, es decir, la forma como el aparato vocal humano puede producir las alteraciones acústicas aéreas y la forma como el sistema auditivo puede percibirlas. En segundo lugar, debe admitirse que los sonidos del habla son elementos de un sistema fonológico peculiar para cada lengua. Todo sistema fonológico limita la manera en que las diversas palabras de la lengua pueden diferir y controla, en parte, el modo en que puede variar su pronunciación.

Durante el habla, una corriente de aire procedente de los pulmones pasa por la laringe, o aparato de la voz, a la garganta y de allí sale a través de la boca. Si la úvula (la campanilla que pende al fondo del paladar) está baja, la corriente de aire sale también por la nariz; si está levantada, los conductos nasales quedan bloqueados. La corriente puede interrumpirse asimismo cerrando los labios, apretando la lengua contra el paladar o cerrando la glotis, un órgano que consta de dos pliegues paralelos de tejido blando (las cuerdas vocales) situados dentro de la laringe.

La corriente de aire puede potenciar

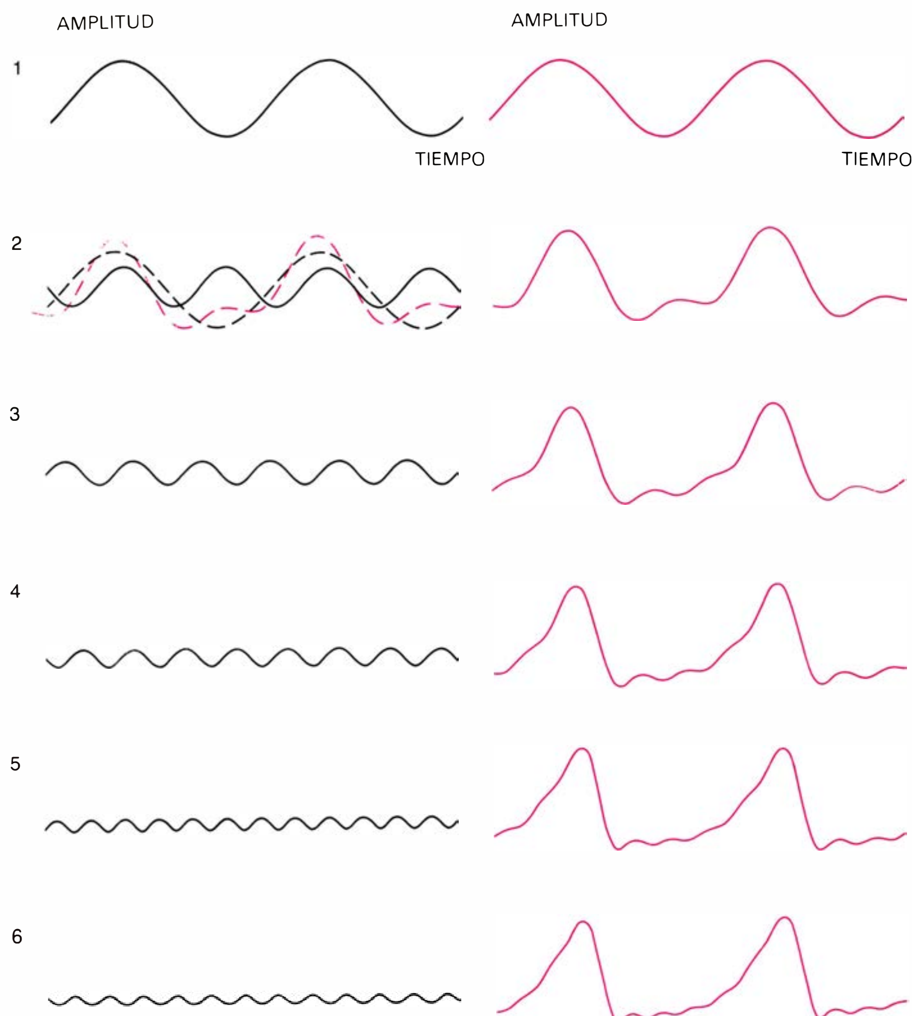
el sonido de tres maneras básicas a lo largo del conducto vocal. La primera, haciendo vibrar las cuerdas vocales de una forma más o menos igual a la doble lengüeta de un oboe o un bajón. Cuando las cuerdas vocales se unen, detienen el paso del aire procedente de los pulmones, con lo que aumenta la presión debajo de ellas. Esta presión provoca la separación de las cuerdas, pero entonces la velocidad del aire que se precipita hacia el exterior reduce la presión en el espacio intermedio. El descenso de presión y la elasticidad de los tejidos vuelven a unir las cuerdas vocales y con ello se inicia de nuevo la presión. La rapidez con que se repite este ciclo constituye la frecuencia fundamental de la voz, que se manifiesta acústicamente a base del tono.

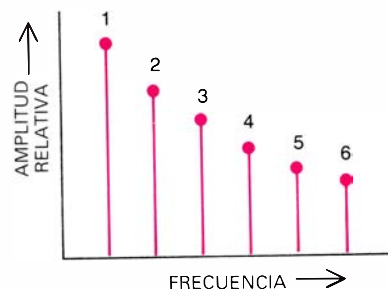
El segundo medio para generar sonido en los conductos vocales consiste en formar una constricción en el canal suficientemente estrecha para causar una turbulencia. Por ejemplo, empujando el aire por una leve separación entre los dientes superiores y el labio inferior se causa una turbulencia que se percibe como el sonido "f". En contraste con los sonidos periódicos que crea la vibración de las cuerdas vocales, los sonidos generados por la turbulencia son aperiódicos, semejantes a un ruido. Cabe incluso la posibilidad de formar sonidos periódicos y aperiódicos al mismo tiempo. En efecto, combinando la vibración de las cuerdas vocales con el ruido de una "f" se produce un sonido resultante que equivale a una "v" [inglesa o francesa].

Un tercer tipo de emisión fónica tiene lugar cuando se libera bruscamente la presión acumulada tras alguna obstrucción. Estos estallidos de energía acústica aparecen en la pronunciación de consonantes como "p", "t" y "k".

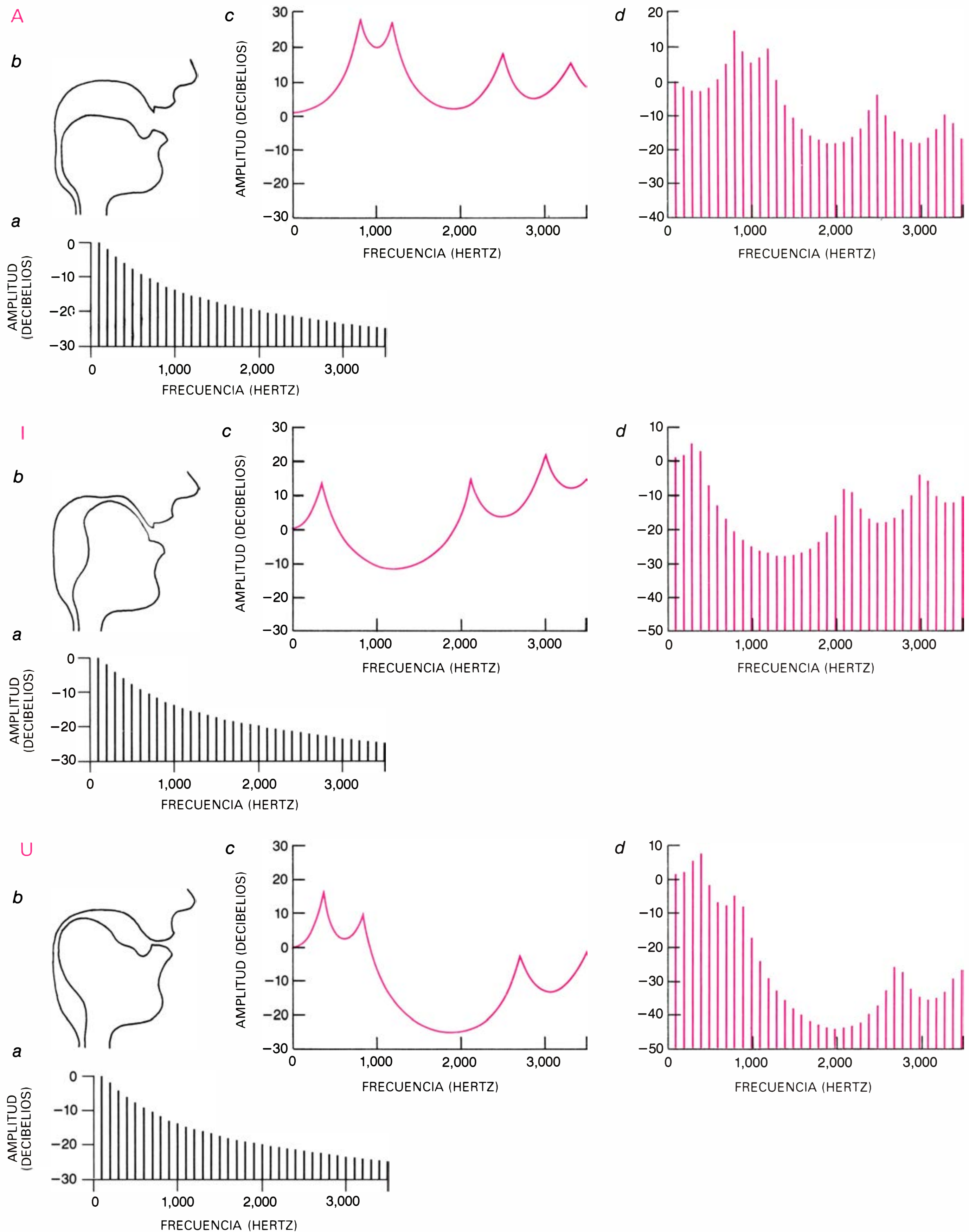
Las tres fuentes de sonido que acabamos de examinar se conforman acústicamente en virtud de la disposición cambiante del conducto vocal. Si de alguna manera las vibraciones de las cuerdas vocales fueran directamente liberadas al exterior, sin pasar antes por la garganta, la boca y la nariz, sonarían como un zumbido, en nada semejante al habla. En cambio, al pasar por la garganta, la boca y la nariz, la cualidad del zumbido cambia profundamente. Gracias, pues, a la forma del conducto vocal, es decir, a las posiciones de la laringe, la lengua, los labios y el velo, se distingue, pongamos por caso, el sonido "e" de "té" del sonido "u" de "tú".

Un medio para comprender esta transformación acústica consiste en el procedimiento matemático conocido por el análisis de Fourier. En 1822, el matemático francés Jean Baptiste Joseph Fourier demostró que toda onda periódica puede representarse como la suma de una serie infinita de sinusoides. Una onda periódica es la que se repite en intervalos uniformes. Si el intervalo de repetición es de t segundos, la frecuencia fundamental de la onda es de $1/t$ hertz. En las series de Fourier para una onda periódica, las frecuencias del componente sinusoidal son armónicos, o múltiplos enteros, de la frecuencia fundamental de la onda en cuestión, a los que hay que asignar amplitudes y fases apropiadas. El transforme o transformación de Fourier es una generalización de la serie de Fourier, pues permite el análisis de ondas ape-





EL PRINCIPIO DE LA SUPERPOSICION refleja la variación temporal en la presión del sonido de la señal que se ha de representar por medio de un espectro de amplitud de sonido, o energía, en diferentes frecuencias. El espectro de amplitud suele ser un medio más útil para suministrar información acústica. La onda (aquí, una onda glotal) puede recibir un tratamiento matemático como una pauta que se repite indefinidamente en el pasado y en el futuro en una frecuencia fundamental. Como lo demostró Fourier, toda onda de este tipo puede descomponerse en una serie de sinusoides múltiplos enteros de la frecuencia fundamental, con diversas amplitudes y fases. Cuando los sinusoides se combinan en cada punto sumando las amplitudes, el resultado equivale a la onda original. Al trazar la amplitud de cada sinusoide de la descomposición en función de su frecuencia, el resultado se convierte en un espectro de amplitud.



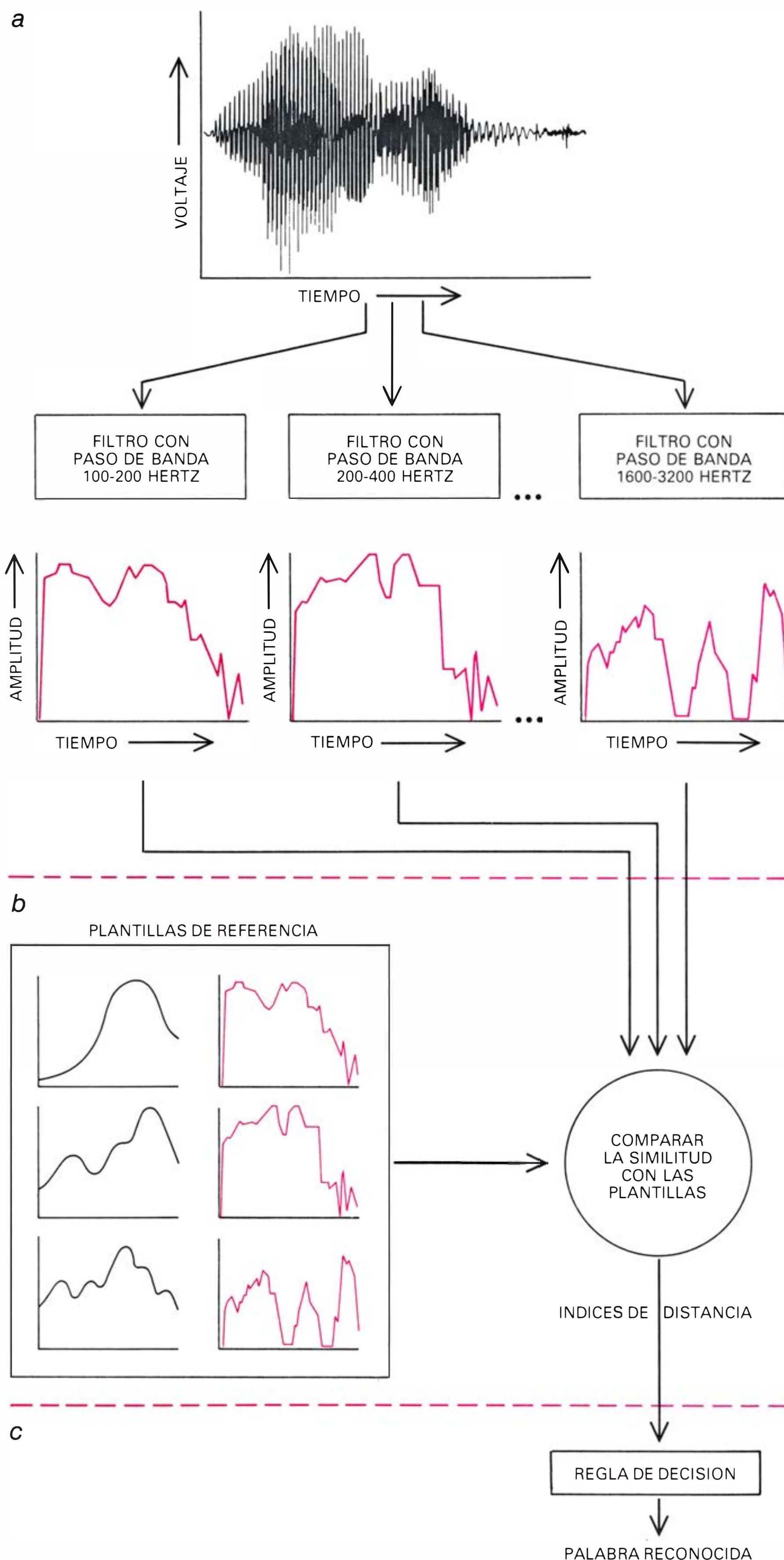
LOS SONIDOS VOCALICOS derivan de diversas configuraciones de la boca, los labios, la lengua y el velo (o paladar blando). La forma resultante del conducto vocal puede concebirse como una serie de cavidades de resonancia que aumentan la energía en ciertas frecuencias y la disminuyen en otras, siguiendo unas pautas predictibles. Estas características de respuesta a unos filtros pueden representarse por una función de transferencia (*c*) para cada posición del modelo del conducto vocal (*b*). Cuando la energía sónica de entrada es periódica, tanto el espectro de entrada (*a*) como el de salida (*d*) son espectros de líneas. En un espectro de líneas, la energía sónica se concentra en armónicos, o múltiplos enteros, de la frecuencia de las cuerdas vocales. Una fuente de sonido aperiódica como la de una vocal cuchicheada no presenta líneas discretas en el espectro, pero la forma del espectro de salida todavía se corresponde con el de la función de transferencia. En la figura aparecen las configuraciones modélicas del conducto vocal para las vocales "i", "a" y "u".

riódicas. Así, el siseo ruidoso del sonido de "f" puede representarse como una suma de componentes sinusoidales a lo largo del continuo frecuencial.

El método más claro para representar ondas fónicas consiste en trazar la variación de la presión del aire en el tiempo. El resultado de Fourier implica que puede también expresarse la misma información mediante un gráfico que muestre la amplitud y la fase en función de la frecuencia de los componentes sinusoidales. Como las diferencias de fase apenas tienen trascendencia perceptual, en la práctica, un sonido de habla puede representarse por su espectro de amplitud, es decir, por un gráfico que indique la amplitud del componente sinusoidal en cada frecuencia.

¿Qué efecto acústico ejerce la forma del conducto vocal sobre el sonido emitido? Cuando se representan los sonidos a partir de sus espectros de amplitud, los efectos son claros [*véase la ilustración de la página precedente*]. El conducto vocal actúa como un filtro sobre el espectro de la fuente fónica, intensificando algunas frecuencias y disminuyendo otras. La filtración selectiva puede también describirse a base de una expresión matemática llamada función de transferencia, de modo que cada una de estas funciones se asocia a cada una de las posiciones que adquieren los órganos del conducto vocal. La función de transferencia suele tener varias cimas frecuenciales, denominadas formantes, en las que se concentra la mayor parte de la energía procedente de la fuente de sonido.

Actualmente, cabe ya la posibilidad de determinar con cierta precisión a qué se debe la dificultad que encuentra el ordenador para traducir de sonidos a palabras, esto es, para pasar de la caracterización acústica de una emisión a la caracterización lingüística del mensaje propuesto. Una de las fuentes de dificultad se debe a que los órganos del habla no asumen una serie de configuraciones fijas en consonancia con las unidades del mensaje. En lugar de ello, las distintas partes del conducto vocal se mueven continuamente siguiendo suaves trayectorias. Algunos investigadores opinan que estos movimientos "discurren" por una serie de posiciones previstas y determinadas por unidades lingüísticas como consonantes y vocales. Otros piensan que aun la más simple de las unidades lingüísticas es inherentemente dinámica. En todo caso, el resultado consta de un movimiento complejo y continuo, que pasa al soni-



METODO CORRIENTE para el reconocimiento de la palabra. Se sirve de los principios del reconocimiento de pautas para distinguir entre pautas acústicas. Se mide y analiza (*a*) la onda del habla, en este caso a través de filtros que dividen la señal en bandas frecuenciales, cada una de ellas con una amplitud de una octava. La salida de cada filtro es la energía que corresponde a su banda. Las salidas se comparan con las plantillas de referencia almacenadas, con lo que reciben unos índices de distancia con respecto a cada plantilla (*b*). Un procedimiento decisorio clasifica el enunciado de entrada a tenor de esos índices (*c*).

do emitido en forma de un espectro de amplitud que cambia constantemente. Estas pautas de cualidad fónica cambiante pueden representarse de un modo conveniente a través de un espectrograma de sonido, esto es, un gráfico donde el tiempo transcurre de izquierda a derecha, la frecuencia aumenta de abajo arriba y la amplitud oscila según una matización que va del gris al negro.

Los movimientos del conducto vocal correspondientes a unidades lingüísticas suelen intersecarse y combinarse con sus vecinos. Así, por ejemplo, al decir "su", la labialización de la vocal "u" precede normalmente al movimiento lingual de la consonante "s". De ahí que los efectos acústicos de los dos movimientos se combinen desde el principio mismo de la palabra. En el habla fluida, esta amalgama se produce también entre una palabra y la siguiente. Los efectos son, a veces, bien claros al oído. Cuando a la "n" de "un" sigue un sonido palatal, como "ch", "n" se palataliza hasta convertirse casi en "ñ": así, "un chico" suena como si fuese "uñ chico", "mancha" como "mañcha", etcétera.

Hay variaciones en el sonido de una palabra que derivan de su posición dentro de la frase, el grado de énfasis y la velocidad de pronunciación. El tamaño y la forma del conducto vocal varía, además, de un individuo a otro, y los hábitos del habla difieren ampliamente según la edad, el sexo, la región geográfica y la extracción social. Por lo demás, la señal que llega a un dispositivo de reconocimiento del habla se ve afectada por diversas circunstancias, al margen de los sonidos emitidos por el hablante, como las condiciones acústicas de la sala, el ruido de fondo y las características del canal de transmisión.

Por estas razones, es difícil dividir una señal de habla en porciones que se correspondan con los elementos del mensaje transmitido por dicha señal, así como traducir segmentos de información acústica en información acerca de la identidad de los fragmentos del mensaje. La gente no encuentra dificultades en comprender el habla, lo que indica que la información requerida ha de estar presente en la señal. La cuestión está en hallarla.

Un punto de partida natural para emprender esta tarea está en el reconocimiento de las palabras. Las palabras suelen ser distintas entre sí en tanto que elementos de un sistema lingüístico y constituyen modelos naturales relativamente estables para un sistema automático de reconocimiento del habla. Y aunque el habla es más que una simple secuencia de palabras, es como mínimo una tal secuencia, de modo que una de las funciones cruciales de un reconocedor del habla consistirá en identificar palabras. A la postre, si un sistema de reconocimiento puede, efectivamente, reconocer palabras con precisión, tendrá éxito; si no, se malogrará.

La mayoría de reconocedores de habla actualmente en uso no son capaces de reconocer palabras en el habla fluida. En lugar de ello, la operación se lleva a cabo sobre palabras aisladas mediante un proceso de reconocimiento de modelos acústicos. Por lo general, el operador debe "entrenar" la máquina introduciendo por medio de un micrófono todas las palabras destinadas al reconocimiento. En algunos casos, el entrenamiento se limita a una sola enunciación de cada palabra por una parte de los hablantes que utilizarán el sistema. En otros casos, todo usuario potencial debe pronunciar varias veces cada palabra. El resultado de este proceso de entrenamiento consistirá en formar un conjunto de "plantillas" coleccionadas que representen modelos acústicos típicos para cada una de las palabras del vocabulario.

Cuando una palabra se presenta al reconocimiento, la máquina analiza la señal acústica, compara los resultados del análisis con las plantillas almacenadas y decide cuál de estas plantillas se parece más a la palabra pronunciada. La máquina puede enumerar asimismo otras posibles compulsas por orden decreciente de similitud. Una vez realizada la clasificación, puede responder al enunciado del operador o emitir una señal adecuada a algún otro artilugio. Cada etapa del procedimiento de compulsa de plantillas (análisis de la señal de habla, comparación con cada plantilla y clasificación de la señal) puede llevarse a cabo por medio de una diversidad de técnicas.

La finalidad de todos los métodos de análisis de la señal hablada consiste en caracterizar la variación temporal de su espectro de amplitud. Acaso el método más simple para evaluar el espectro sea el del cálculo de cruces sobre cero. Este método consiste en contar el número de veces que el voltaje análogo de la señal de habla cambia de signo algebraico (de más a menos o de menos a más) en un intervalo dado. El número de estos cruces axiales guarda relación con la frecuencia.

Para mejorar este método del cruce sobre el valor cero se ha procedido a filtrar la señal del habla en tres bandas frecuenciales. Los cruces sobre cero se miden por separado en cada banda con el fin de proporcionar estimaciones aproximadas sobre las frecuencias de los tres primeros formantes. Estas medidas son útiles para clasificar los sonidos vocálicos e incluso suficientes para discriminar palabras bien distinguidas en un vocabulario reducido. El método de los cruces sobre cero resulta económicamente atractivo porque puede llevarse a cabo mediante dispositivos electrónicos sencillos.

Un procedimiento más elaborado para la evaluación espectral es el del banco de filtros. La señal del habla queda escindida por filtración en unas 20 o 30 bandas frecuenciales a lo largo de la gama de frecuencias típicas del habla humana. La producción de salida de cada filtro equivale a la medida de la energía en la respectiva banda frecuencial. Los niveles de energía se vuelven así idóneos para la comparación directa con los de una plantilla. La Transformación Rápida de Fourier facilita un método general y computacionalmente eficiente para evaluar el espectro de amplitud de una señal a partir de su onda en desarrollo temporal. Este algoritmo proporciona uno de los diversos modos de obtener información del banco de filtros en forma exclusivamente digital.

Más recientemente se ha introducido un nuevo método, llamado de análisis predictivo lineal, para evaluar el espectro de amplitud del habla. En realidad, los estadísticos lo han ido empleando durante algún tiempo con el nombre de análisis autorregresivo. El método pronostica la amplitud de una onda habla-

METODOS DE EVALUACION de los espectros de amplitud en intervalos cortos de una palabra (aquí, la palabra "language" ["lenguaje"]); tratan de destacar la información lingüística pertinente de una manera computacionalmente efectiva. Los cálculos de cruces sobre cero aprovechan el hecho de que, a medida que aumenta la frecuencia, aumenta también el número de veces que el voltaje análogo de la señal acústica cambia de signo. En el método de bandas filtradas la señal se divide en varias bandas frecuenciales y se mide la cantidad de energía en cada banda. Estas mediciones producen un espectro de amplitud para el intervalo respectivo. La Transformación Rápida de Fourier es un algoritmo general y computacionalmente eficiente para valorar el espectro de amplitud de la señal a partir de su onda desarrollada en el tiempo. Es una de las diversas maneras que hay de computar la información del banco de filtros en forma digital. La apariencia rugosa del espectro está causada por los armónicos de tono o por otra estructura fina del espectro. El cuarto método de evaluación espectral, análisis predictivo lineal, emplea un modelo del conducto vocal para generar espectros frecuenciales sucesivos. Su ventaja radica en que genera un espectro suave y continuo para cada muestra. Los espectros en color oscuro están construidos a partir del mismo intervalo de la señal que varía con el tiempo. Existen otros métodos de evaluación espectral.

1

VOLTAJE

TIEMPO

FILTROS CON PASO DE BANDA

200-900 HERTZ

800-2500 HERTZ

2000-3200 HERTZ

NUMERO DE CRUCES SOBRE CERO

25 20 15 10 5

60 50 40 30 20

75 65 55 45

TIEMPO

EVALUACION APROXIMADA DE LAS FRECUENCIAS DOMINANTES EN LAS TRES BANDAS

2

VOLTAJE

TIEMPO

FILTROS CON PASO DE BANDA

200-252 HERTZ

252-317 HERTZ

3200-4032 HERTZ

e_1 e_2 e_n

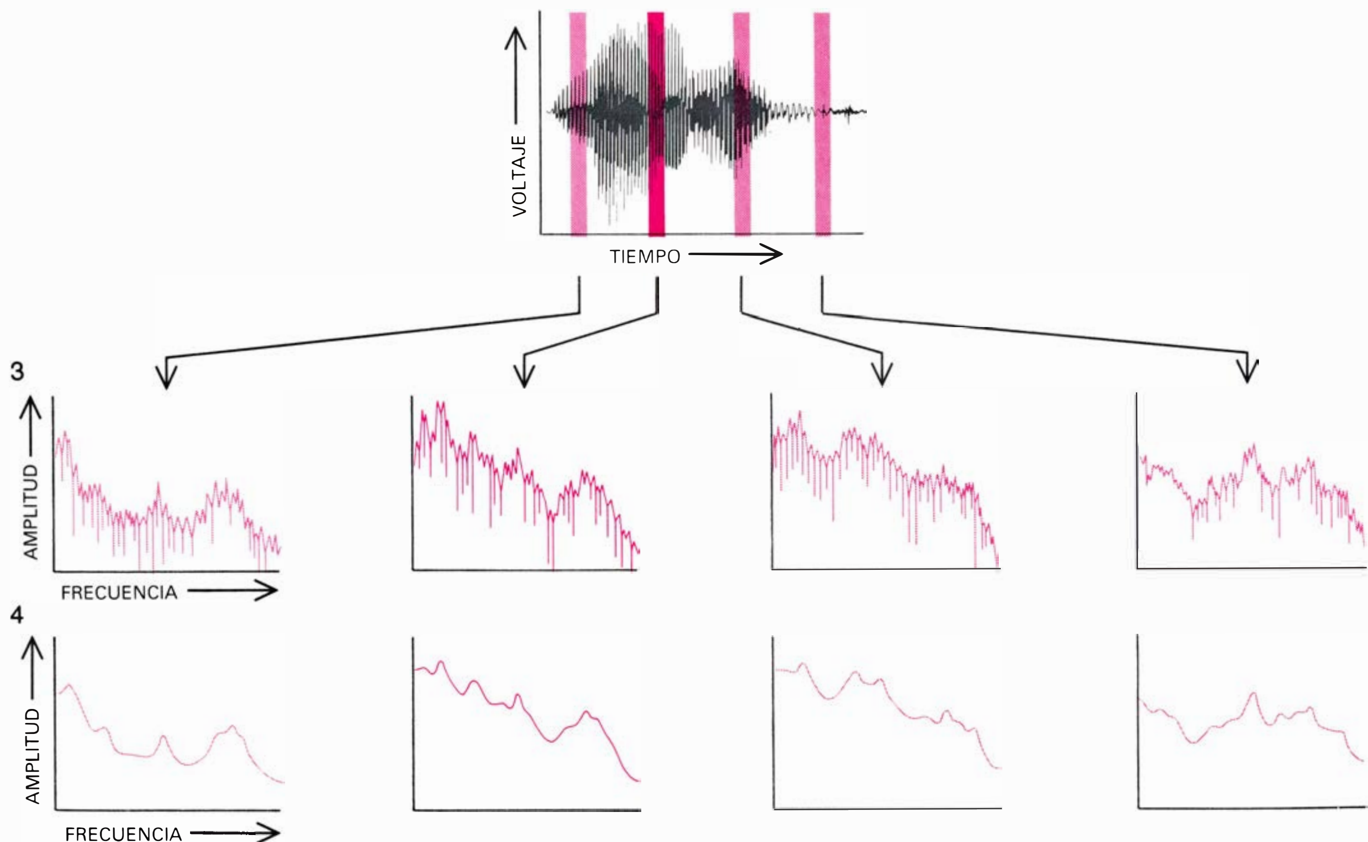
AMPLITUD

TIEMPO

AMPLITUD

e_2 e_1 e_n

FRECUENCIA

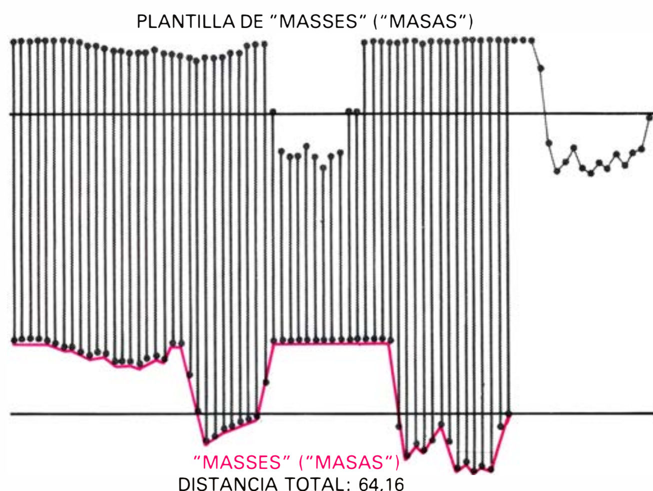


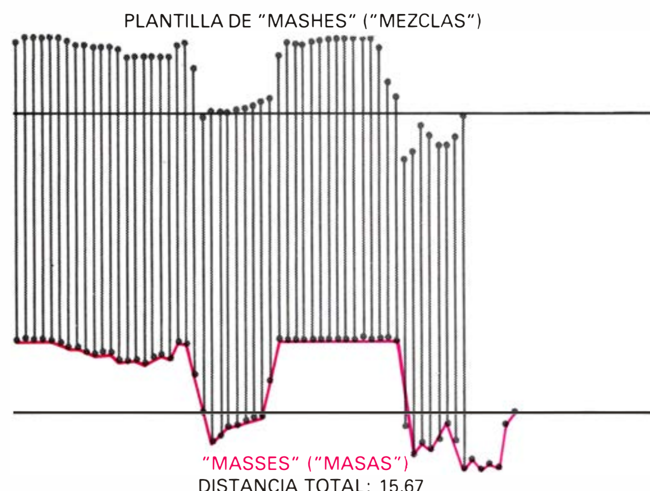
da, en un instante dado, a partir de una suma ponderada (o combinación lineal) de sus amplitudes en un pequeño número de instantes iniciales. Los coeficientes, o ponderaciones, que proporcionan la mejor evaluación de la onda perteneciente al habla auténtica pueden, entonces, convertirse matemáticamente en una evaluación del espectro de la amplitud, pues el examen detallado del análisis predictivo lineal del habla es especialmente adecuado desde el momento en que equivale matemáticamente a tratar el conducto vocal como si fuese un tubo de diámetros variables o, en otros términos, como una secuencia de cavidades de resonancia. El modelo es bien fidedigno para el habla sonora no nasalizada. Como se trata de un modelo de las resonancias del conducto vocal y no de la vibración de las cuerdas vocales, el espectro de predicción lineal resulta uniforme. En él no se destaca ninguno de los armónicos tonales y, en consecuencia, la estructura de formantes de la onda hablada, tan importante para el reconocimiento del habla, sobresale con claridad.

Durante la etapa de comparación o compulsa de plantillas, puede aprovecharse la estructura fonológica de una palabra de un modo indirecto. Una palabra hablada consta de una secuencia de gestos vocales, que da lugar a una pauta fónica variable en el tiempo. Las partes de la pauta fónica raramente tienen la misma duración en distintas enunciaciones de la misma palabra, pero su secuencia es mucho más constante. Por ejemplo, la palabra "fábula" empieza con un ruido de "f", seguido por una pauta de formantes en transición correspondiente a la abertura labial, que alcanza su máximo, hasta un nuevo cierre, casi completo, para la "b", mientras la lengua adopta la posición plana de "a"; tras la constricción labial con redondeamiento de "b", se produce otra abertura brusca con retracción posterior de la lengua, típica de "u"; finalmente, viene otra pauta de movimiento espectral con alisamiento de labios y elevación del ápice de la lengua para "l", seguido de un nuevo aplanamiento lingual, propio de "a". La distribución temporal de estas pautas puede variar considerablemente en distintas emisiones, pero todas han de presentarse en el orden descrito si la enunciación ha de contar como una razonable ejecución de la palabra "fábula".

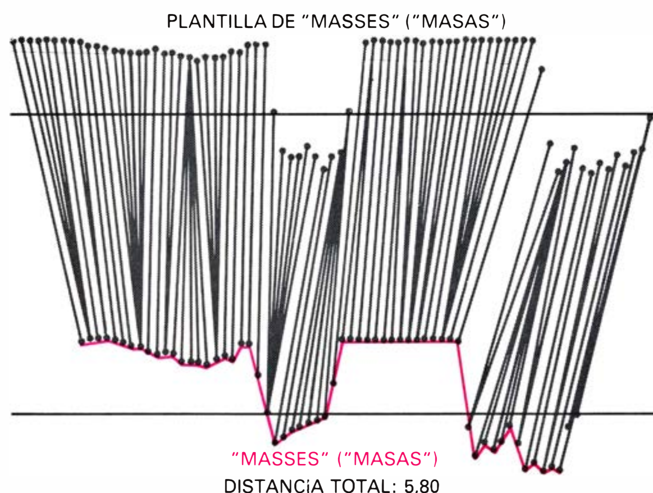
A causa de las diferencias en la distribución temporal, las distintas partes de la palabra pueden quedar muy lejos de las previsiones establecidas en la co-

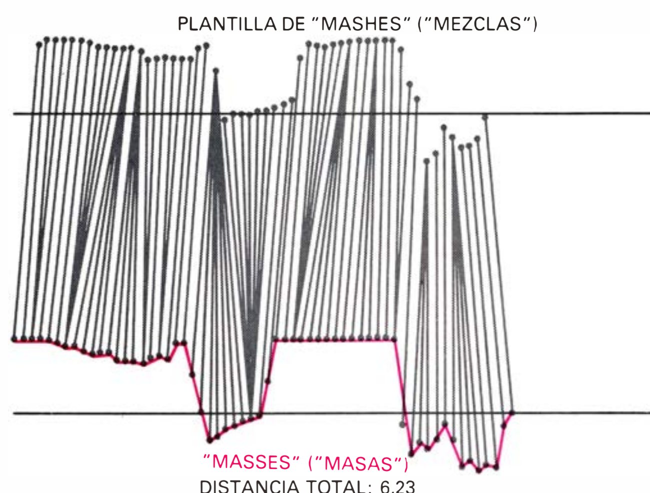
COMPULSA DIRECTA





COMPULSA POR MEDIO DE PROGRAMACION DINAMICA



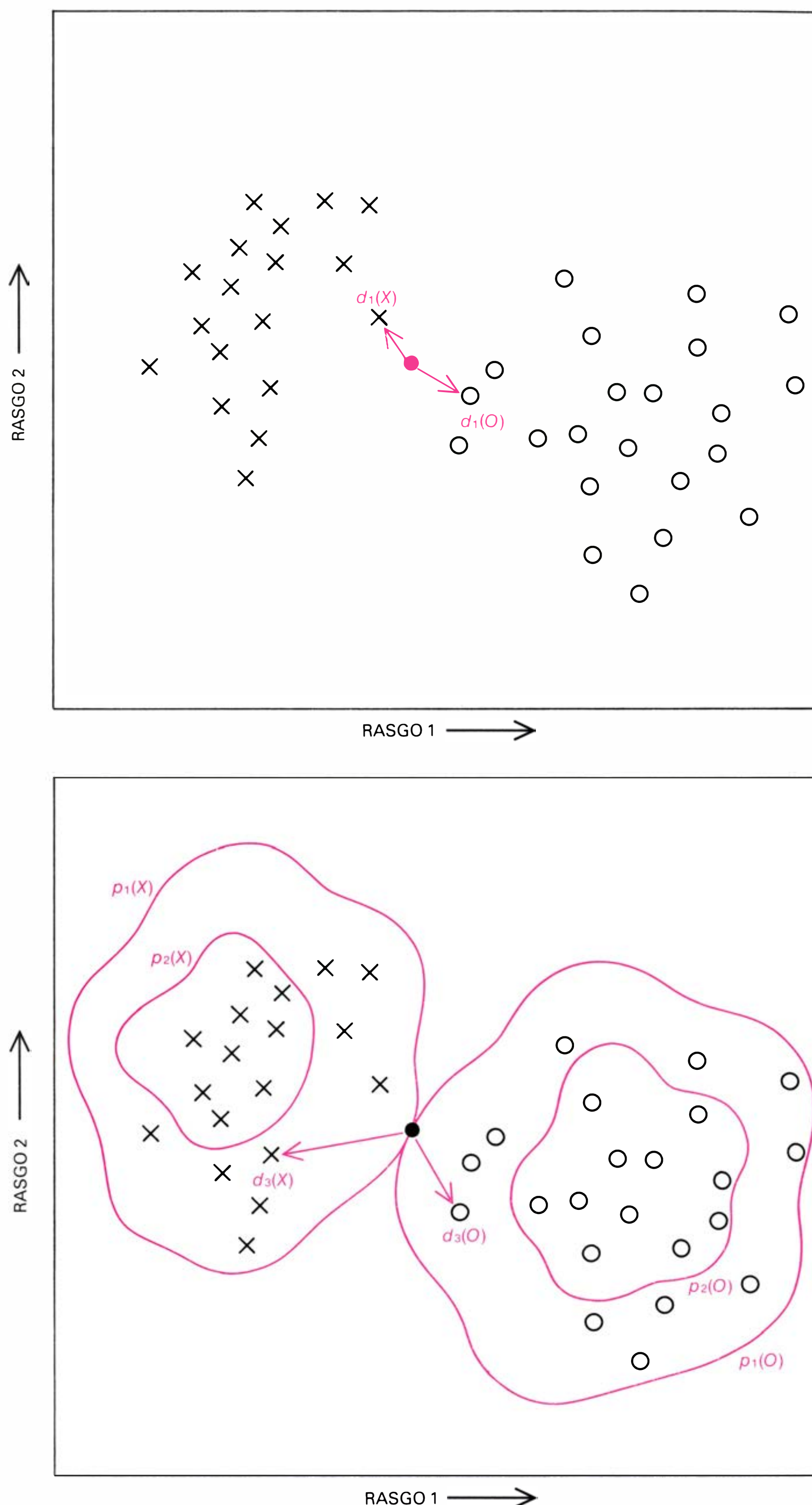


ETAPA DE LA COMPARACION en el reconocimiento de la palabra; se realiza comprimiendo y ampliando plantillas almacenadas de acuerdo con un proceso de modelación llamado de programación dinámica. En cada plantilla, la programación dinámica intenta asociar cada uno de los armazones (frames) de la palabra introducida con algún armazón de la plantilla de modo que encuentre la medida mínima de distancia en la correspondencia total entre la entrada y dicha plantilla. La alineación temporal no uniforme de la plantilla almacenada con la palabra hablada permite cierta variación en el ritmo del habla y en la longitud relativa de las vocales y consonantes de una palabra. La compulsa de las plantillas (*negro*) con la entrada (*color*) sin programación dinámica produce una falsa identificación, indicada por los resultados de distancia, que se corrige al aplicar el procedimiento de compresión y expansión. La programación dinámica se realiza a menudo con ordenador, pero no debe confundirse en absoluto con la programación misma del ordenador.

rrespondiente plantilla. Como el orden de los eventos es bastante constante, la desviación puede corregirse ensanchando la plantilla en diversos puntos y comprimiéndola en otros, hasta encontrar una compulsa matemáticamente óptima. La alineación temporal no uniforme se consigue por medio de un procedimiento, denominado de programación dinámica y desarrollado por Richard E. Bellman, de la Facultad de Medicina en la Universidad de Southern California, para resolver problemas en el diseño de servomecanismos. Se trata de una técnica para la modelación matemática que a menudo se lleva a cabo con ayuda de ordenador y que no debe confundirse con la programación misma del ordenador.

La comparación comporta una estimación del grado de similitud entre el sonido de entrada y el sonido representado por la plantilla almacenada. En definitiva, este procedimiento común a todos los reconocedores del habla consiste en una estrategia decisoria que suele basarse en una medida estadística de la aproximación al ajuste entre el enunciado de entrada y una plantilla cualquiera. A cada plantilla se le asigna un punto en un espacio abstracto, de modo que la posición de dicho punto quede definida por las características espectrales de la plantilla en cuestión. El enunciado que ha de clasificarse queda, a su vez, representado como un punto en el mismo espacio. El reconocedor calcula, entonces, la distancia que hay en el espacio entre dicho enunciado y cada una de las plantillas. Al fin, elige la plantilla o clase equivalente de plantillas más próxima al enunciado en un sentido estadístico.

El alcance de los sistemas de reconocimiento automático en la identificación de palabras aisladas apenas puede compararse con el de las personas. Incluso para los más potentes reconocedores de palabras, el número de errores crece rápidamente en cuanto se aumenta el vocabulario a más de unos pocos centenares de entradas. El índice de errores es todavía superior en cuanto se incorporan hablantes y condiciones acústicas desconocidas. En un experimento reciente, fueron pronunciadas algunas palabras aisladas, a partir de un vocabulario de 26.000, por parte de una serie de hablantes desconocidos por el oyente. Estas palabras fueron identificadas con un índice de error por debajo del 3 por ciento. Las facultades humanas para el reconocimiento del habla tienen asimismo una notable tolerancia con el ruido de fondo, pues una conversación suele progresar aun en medio de



CLASIFICACION DE UN SONIDO DE ENTRADA. Consiste dicho método en hallar la distancia más corta, en un espacio de rasgos acústicos, entre la entrada (representada por un punto) y una plantilla o clase de plantillas almacenadas (representadas por las *X* y *O*). La estrategia decisoria más simple viene a escoger la plantilla más próxima (*véase el gráfico superior*), con lo que la entrada queda clasificada como un sonido "a" (una *X*). Cuando varias plantillas representan sonidos lingüísticamente equivalentes (supongamos, por ejemplo, que el ordenador tiene que reconocer las voces de diversos hablantes), la estrategia decisoria puede tomar en consideración clases enteras de plantillas. Un método calcula la distancia que hay entre la entrada y el tercer vecino más cercano en cada clase (*véase el gráfico inferior*). En este caso, la entrada queda clasificada como un sonido "o" (una *O*). En ciertas condiciones, cabe la posibilidad de delimitar zonas de igual densidad dentro de las cuales el número de muestras de plantilla por unidad de zona es constante. Puede encontrarse, entonces, la zona de mayor densidad que pasa por la entrada; ahora bien, puesto que p_1 (*X*) es mayor que p_1 (*O*), la entrada se clasifica como un sonido "a" (una *X*).

una ruidosa tertulia. En cambio, ningún sistema existente de reconocimiento automático puede acercarse a un nivel de rendimiento semejante.

En las pruebas realizadas hasta ahora para reconocer el habla continua, la disparidad entre la actuación humana y la del ordenador es todavía más evidente. Mientras la gente suele encontrar más fácil reconocer las palabras en medio de un contexto, para un sistema automático el reconocimiento del habla fluida es muchísimo más difícil que el de las palabras aisladas. Uno de los problemas cruciales es el de la coarticulación, que provoca la mezcla de palabras en sus respectivas lindes y hace muy complejas e inestables las pautas espectrales sujetas a reconocimiento. En el habla continua, no hay signos acústicos claros que denuncien las fronteras entre palabras, con lo que la compulsa directa de las plantillas se hace extremadamente difícil. En esencia, cada plantilla requiere un cotejo con todos los posibles intervalos de la enunciación por medio de una variante del método de programación dinámica.

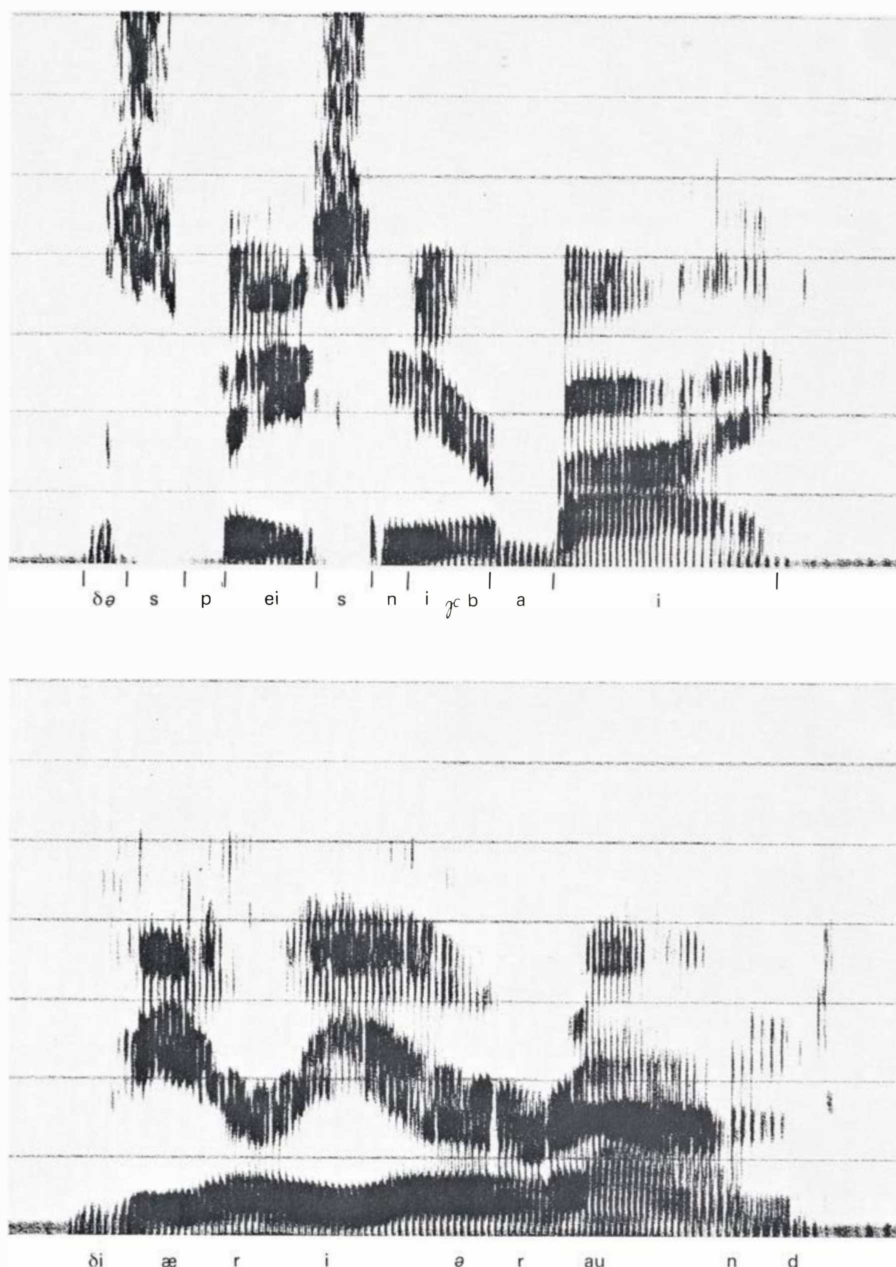
La tarea computacional se reduce un tanto si se introduce el requisito de que los intervalos sean contiguos, de manera que el final de una palabra se encuentre con el principio de la siguiente. Aun así, la complejidad combinatoria del proceso aumenta demasiado de prisa para que se considere una solución práctica al problema general de reconocer el habla continua. La compulsa directa de plantillas sólo puede resultar provechosa cuando se reduce el ámbito de enunciados posibles. Con la tecnología actual, el procedimiento puede funcionar durante el tiempo real (es decir, a medida que se emite el enunciado) si las secuencias no sobrepasan las cinco palabras, tomadas de un vocabulario de unas 20 entradas.

En vez de buscar cada una de las posibles pautas en todos los lugares de la señal, un sistema de reconocimiento del habla continua busca unidades lingüísticas de un modo más estricto, siguiendo la secuencia desde el principio del enunciado hasta el final. La señal hablada queda dividida en intervalos que se corresponden con pautas acústicas específicas; los intervalos, a su vez, se clasifican de manera que satisfagan, en lo posible, las categorías de un mensaje lingüístico potencial. A estas técnicas las llamaremos de segmentación y rotulación. Son procesos que pueden llevarse a cabo de muchos modos, pues los intervalos buscados pueden corresponder a palabras enteras o a unidades lingüísticas más pequeñas, como sílabas, pares fonemáticos o fonemas.

El procedimiento más sencillo para segmentar y rotular automáticamente consiste en obligar al hablante a hacer una breve pausa entre las palabras. Así, las pausas, que aparecen como intervalos de baja energía fónica, constituyen una indicación fiable de fronteras de palabras. Una vez segmentadas las palabras, puede procederse al análisis respectivo. Ahora bien, aunque este método funciona bien, en realidad no enfoca la cuestión de reconocer el habla fluida. Hay, sin embargo, otros métodos a nuestro alcance.

Las interrupciones del espectro, los altibajos de energía de ciertas bandas frecuenciales y otros signos acústicos facilitan claves sobre eventos articulatorios: el cierre o abertura del conducto vocal o el comienzo o final de la vibración laríngea. Esto sugiere que la segmentación y la rotulación pueden llevarse a cabo a partir de las unidades fonológicas básicas de que constan las palabras.

La mezcla y difuminación de información acústica a través de los límites respectivos afectan a la forma acústica de las unidades más pequeñas del habla aún más que a la de las palabras. Por ello es difícil identificar estas unidades



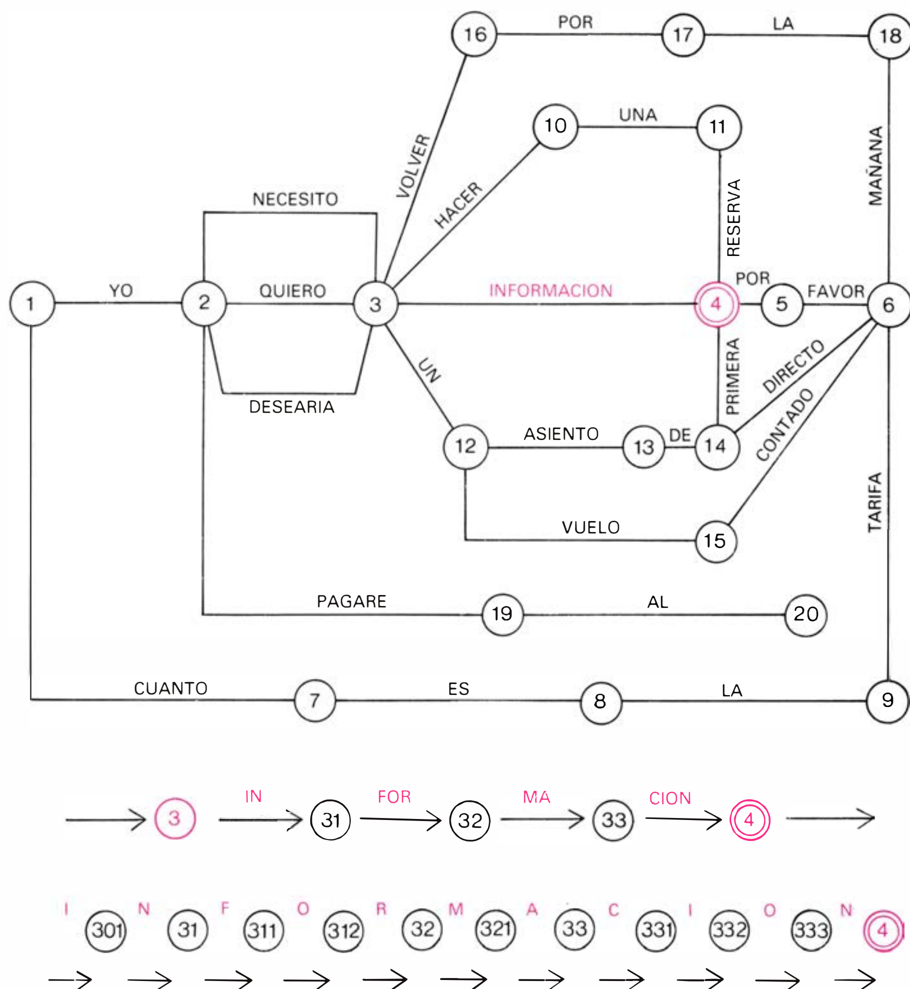
RESULTA DIFICIL LA SEGMENTACION DE PAUTAS ACUSTICAS EN PALABRAS u otras unidades lingüísticas, a causa de la difuminación temporal de los sonidos del habla. Ciertos sonidos, no obstante, ofrecen más discontinuidad espectral que otros. La alternancia de consonantes y vocales en la frase "The space nearby" [ðə'speis'niə,bai] ("El espacio inmediato") se presenta a base de discontinuidades relativamente claras. En cambio, en la ilustración inferior, la suave secuencia de vocales fundidas en la frase "The area around" [ði'æʳiəʳ'ɑound] ("La zona de alrededor") dificulta la segmentación.

mediante la compulsa de plantillas, pues los errores de segmentación serán al menos tan frecuentes para las unidades más pequeñas del habla como para las palabras. A pesar de todo, puede haber razones que apoyen la segmentación en unidades más pequeñas a medida que aumentan los vocabularios de los reconocedores del habla.

Hay en inglés unas 300.000 palabras, muchísimas más de las que cabe confrontar para la compulsa de plantillas. Además, es difícil prever los efectos de la mezcla en las fronteras cuando se emplean plantillas de palabra. Las sílabas en inglés llegan a unas 20.000, todavía demasiadas para su identificación fácil y fidedigna. Por otro lado, los efectos de la mezcla en las fronteras son todavía más perjudiciales para la compulsa de plantillas con sílabas que con palabras. En cambio, hay sólo unos 40 fonemas en inglés (esto es, elementos lingüísticos básicos, como consonantes y vocales), los cuales pueden, a su vez, descomponerse en una docena de rasgos fonológicos que establecen las características distintivas de la forma del conducto vocal y el control de la laringe. Estos rasgos pueden también combinarse directamente en unidades de tipo silábico. No obstante, a medida que decrece el número de unidades lingüísticas, su relación con las pautas de sonido se vuelven más abstractas, más complejas y peor comprendidas. La segmentación y rotulación de estas unidades tan pequeñas de habla mediante las técnicas hoy asequibles conduce a altos índices de error. Y, sin embargo, si las limitaciones que impone el código lingüístico pueden compensar los errores, o si llegan a encontrarse métodos más fiables de análisis, el reducido número de unidades fonológicas básicas brindará una ventaja decisiva en favor de estos elementos fundamentales.

Una dificultad común a todos los procedimientos es que la probabilidad de error es mucho mayor en una serie de clasificaciones independientes que en una sola clasificación. En una frase de tres palabras, aun cuando la probabilidad de reconocer la palabra correcta en cualquier posición dada sea de 0,8, la probabilidad de reconocer correctamente la frase entera apenas pasa de un medio (0,8 × 0,8 × 0,8).

Un modo de contrarrestar este inconveniente consiste en incorporar restricciones impuestas por el propio código, tales como limitarse a las secuencias permisibles de palabras en una oración o de sílabas en una palabra. Una rama de la matemática, llamada teoría del lenguaje formal, facilita varios métodos



UNA GRAMATICA DE ESTADOS FINITOS constituye, computacionalmente hablando, el medio más simple de imponer constricciones sintácticas (o de orden de palabras) en el reconocimiento de oraciones. La gramática del presente esquema forzaría el ordenador a clasificar toda secuencia de palabras acústicamente posibles como una de las 26 oraciones localizables a través del diagrama de estados, empezando por el estado 1 y terminando por el estado 5 o 6. Por ejemplo, una oración posible sería "Desearía un asiento de primera clase, por favor". Los principios del diagrama pueden ampliarse asimismo a niveles de análisis más bajos que el de palabra, como el silábico y fonético de la palabra "información". Las gramáticas de los sistemas para el reconocimiento experimental admiten miles de millones de oraciones.

para especificar y utilizar dichas restricciones. Aplicando algunos de los principios básicos de la teoría del lenguaje formal cabe la posibilidad de establecer descripciones precisas y eficientes, o gramáticas formales, de las secuencias lingüísticamente posibles de sonidos y palabras. Pueden incluso establecerse programas de ordenador que empleen estas gramáticas para reconocer secuencias lingüísticas formalmente correctas.

Una manera sencilla de explotar la estructura gramatical recurre a una elaboración matemática llamada diagrama de estados. Un diagrama de estados define cada una de las posibles oraciones que la máquina puede reconocer. Cada trayecto desde el punto de partida del diagrama hasta los puntos terminales representa una oración aceptable. Entonces, a base de mediciones acústicas, el reconocedor asigna una probabilidad a cada transición del diagrama. Puede, en consecuencia, calcularse una probabilidad para cada trayecto formando el producto de las probabilidades de todas las transiciones que componen el trayecto en cuestión. La oración elegida es la que viene representada por el trayecto de más alta probabilidad. Desde luego, esta técnica puede reducir sensiblemente el índice de errores en el reconocimiento de oraciones, ya que es capaz de optar por una palabra con una probabilidad relativamente baja en una posición dada a fin de intensificar la verosimilitud de la transcripción en su conjunto.

Esta reducción en el índice de errores se puso de manifiesto con un sistema de base fonemática para el reconocimiento de japonés fluido, probado en los Laboratorios Bell y en los Nippon Telegraph and Telephone Electrical Communication Lab. La segmentación y rotulación de fonemas resultó correcta sólo el 60 por ciento del tiempo. En cambio, con un tratamiento sintáctico se llegó al 70 por ciento de corrección

en el reconocimiento de oraciones de una longitud media de 25 fonemas. Aunque un 70 por ciento del reconocimiento no es adecuado para una comunicación fehaciente, el resultado es extraordinario a la vista de la escasa probabilidad de encontrar una oración correcta sin tratamiento sintáctico: viene a ser como una posibilidad entre tres millones.

Un diagrama de estados puede, incluso, mejorar la eficiencia del reconocimiento del habla continua mediante una alineación de tiempo no lineal. En lugar de compulsar cada una de las plantillas con cada uno de los intervalos de la oración introducida, el sistema de reconocimiento sólo comprueba aquellas plantillas que se ajustan a las secuencias admisibles descritas por el diagrama de estado. Este procedimiento elimina mucho cálculo superfluo, puesto que sólo puede aparecer un pequeño subconjunto de las palabras del vocabulario en una posición dada de la oración. Un dispositivo que emplee una alineación de tiempo sintácticamente orientado puede reconocer oraciones conexas de más de 20 palabras compuestas a partir de un vocabulario de más de 100 entradas.

Hasta aquí, hemos descrito los símbolos fonológicos que se corresponden con la realidad acústica del habla y su organización gramatical en palabras y frases. Estos símbolos forman el código lingüístico del habla. El propósito del código lingüístico consiste en transmitir mensajes significativos, esto es, información semántica. De ahí que la información semántica imponga nuevas restricciones al modo como pueden combinarse los símbolos de una lengua para formar mensajes.

Una máquina que elabora la información semántica codificada en el habla se propone realizar una tarea mucho más compleja y sutil que una máquina que se limite a reconocer palabras. Para asumir el significado, no sólo ha de reconocer las pautas acústicas, sino que ha de manipular, además, representaciones abstractas de la realidad. En otras palabras, ha de simular al menos ciertos aspectos importantes de la inteligencia humana.

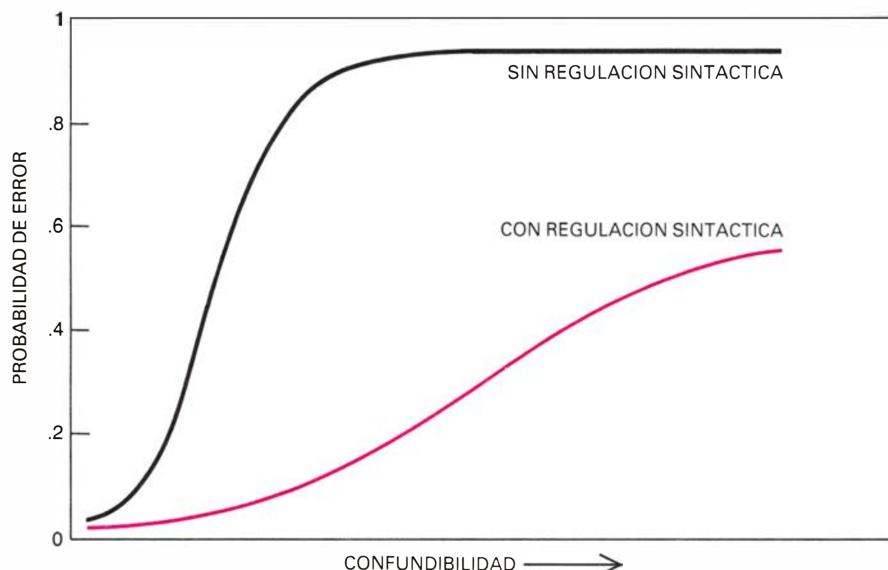
En los Laboratorios Bell hemos incorporado un procesador semántico rudimentario a un sistema diseñado para emular el proceso total de la comunicación humana a través del habla. El operador se comunica por teléfono con el sistema. El ordenador, que en este caso actúa como si se tratara de un vendedor de billetes para una línea aérea, responde a través de voz sintetizada. La integración de las funciones necesarias en un solo dispositivo nos ha permitido estudiar la interacción de los distintos subsistemas y su control.

Como simulación completa de la comunicación humana, la máquina de los Laboratorios Bell constituye el más avanzado sistema que conocemos. Los componentes por separado, sin embargo, son menos avanzados que los empleados en otros laboratorios para experimentos análogos. Existen sistemas de reconocimiento del habla que funcionan con vocabularios muy superiores al de las 127 palabras que reconoce nuestra máquina, y no faltan otros dotados de una sintaxis más flexible que la nuestra. Existen, asimismo, procesadores semánticos más refinados que aceptan entradas mecanografiadas en lugar de habladas. También hay procesadores que responden más de prisa que el nuestro. Una pregunta formulada en 10 segundos recibe una respuesta al cabo de unos 50 segundos en nuestro sistema. Desde luego, esperamos mejorar el rendimiento de cada uno de sus elementos.

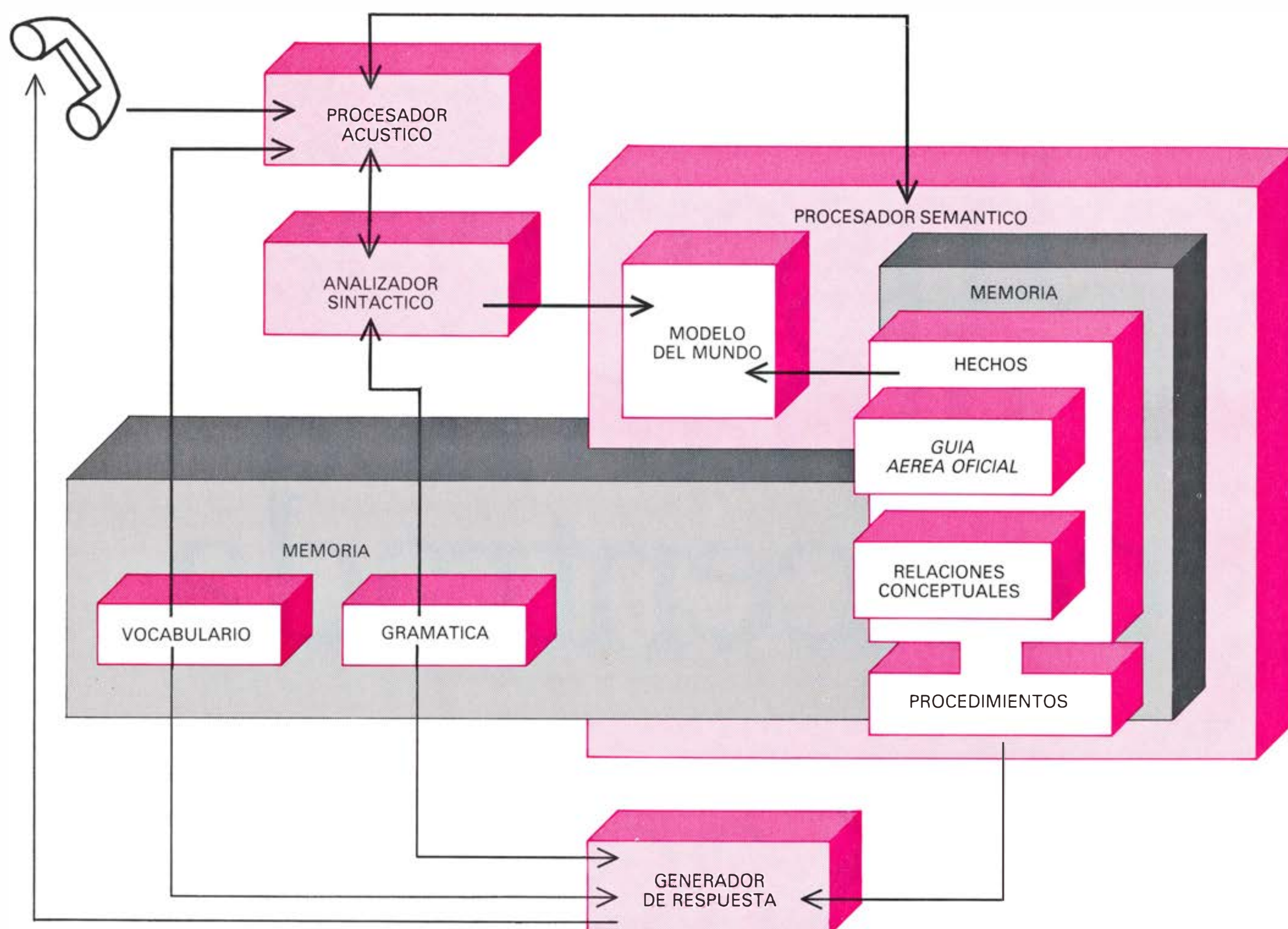
En el sistema para la información aérea se emparejan el procesador acústico y el analizador sintáctico, a fin de que el primero compruebe cada una de las palabras hipotéticas que le envía el segundo para verificar la información espectral. El resto del sistema, salvo dos unidades de memoria que aprovechan todos los componentes, está dedicado a la elaboración semántica.

El procesador semántico contiene un modelo del mundo –cuyo estado puede cambiar a medida que progresa una conversación– y un módulo de memoria que no puede alterarse. El modelo del mundo se basa en un conjunto de conceptos, en donde cada concepto puede adquirir una serie de valores. Entre los conceptos en cuestión se cuentan los de "destino", "día de salida" y "hora de salida". A lo largo de una conversación, estas categorías podrían recibir los valores de "Boston", "martes" y "hora 17", mientras que otro estado del mundo podría corresponder a los valores "Chicago", "desconocido" y "desconocida". El procesador semántico determina un nuevo estado a partir del presente, así como de las palabras de la oración introducida y de las transiciones aparecidas en el diagrama de estados, empleadas al generar la oración. La necesidad de contar con las dos últimas fuentes de información refleja el hecho de que el contenido semántico está en función tanto de las palabras como de sus relaciones en el marco oracional.

Las unidades de memoria almacenan dos tipos de información: hechos y procedimientos. Los hechos, a su vez, se dividen en dos clases. Los horarios de los vuelos quedan almacenados como parte de la *Guía Aérea Oficial,* pero hay que almacenar también las relaciones entre los conceptos de la *Guía.* Si se solicita al sistema el tiempo requerido para un determinado vuelo, puede calcularlo a base de los tiempos especificados de salida y llegada. Ahora bien, para realizar todo esto hay que disponer de las diferencias de tiempo en cada ciudad, según los husos horarios (pues la *Guía Aérea Oficial* sólo opera con el horario local).



LA CONFUNDIBILIDAD de una señal de habla está en función del tamaño del vocabulario de entrada, de la similitud acústica de los elementos que deben distinguirse, del número de hablantes a que debe atender el sistema y de la cantidad de ruido del canal comunicativo. Los errores tienden a volverse más frecuentes a medida que aumenta la confundibilidad. Las constricciones sintácticas pueden reducir este inconveniente. Esta pauta de error es tan válida para los oyentes humanos como para las máquinas.



SIMULACION COMPLETA de la comunicación humana por medio del habla mediante un sistema automático que los autores y sus colegas han construido en los Laboratorios Bell. En el diagrama se muestran las relaciones funcionales entre las partes más importantes del sistema. El operador pide información sobre los horarios aéreos con ayuda del teléfono y el ordenador contesta por medio de voz sintetizada. Las flechas de trazo grueso siguen el curso de la información relativa al reconocimiento del habla. La generación de una respuesta se refleja con flechas de trazo fino. Los módulos de memoria destinados al tratamiento semántico comprenden hechos y procedimientos relativos a vuelos y reservas. La memoria no semántica almacena plantillas del vocabulario y reglas gramaticales utilizadas tanto en el reconocimiento del habla como en su síntesis. El procesador semántico incluye asimismo un modelo del mundo constantemente puesto al día con datos que están basados en las preguntas del operador y en la información de la memoria semántica.

Por su parte, los procedimientos son programas para determinados propósitos que emplean hechos almacenados con el fin de obtener nueva información a partir de datos introducidos y del estado actual del modelo del mundo. Por ejemplo, un programa es un calendario perpetuo que puede encontrar el día de la semana de una fecha cualquiera. La conversión es indispensable porque una pregunta puede especificar tan sólo una fecha de salida, mientras que la *Guía Aérea Oficial* se organiza a base de los días de la semana.

Cuando una instrucción interna exige una réplica del operador, el sistema activa un codificador lingüístico. El analizador semántico dice al codificador qué conceptos hay que comunicar del modelo del mundo. Entonces, el codificador recobra la gramática y el vocabulario de la memoria y convierte los conceptos en una secuencia de símbolos. El sintetizador, entonces, transforma la secuencia en habla.

¿De qué manera puede mejorarse el arte de comprender el habla? Hay dos objetivos básicos a la vista. A corto plazo, es preciso alcanzar una mejor comprensión sobre la delicada estructura de la comunicación hablada. Esto debería comprender una información detallada sobre el tipo de análisis de señal que hace el oído humano, así como un mayor conocimiento acerca de la relación que hay entre los símbolos fónicos (como fonemas y sílabas) y los sonidos reales. Hay que promover medios más eficientes para explotar esta información e incorporarla a los sistemas de reconocimiento.

Para un futuro más remoto, no faltan campos de investigación capaces de aportar significativos avances. Hemos subrayado ya que el código del habla comprende una serie de tipos coexistentes de estructura: fonológica, sintáctica y semántica. Es menester una teoría general de estos códigos complejos, en particular para coordinar y controlar las interacciones entre los niveles. Conviene asimismo conseguir una mejor comprensión de los procesos que experimentan las personas al adquirir la primera lengua. Aunque los actuales reconocedores del habla están "entrenados", el entrenamiento es rudimentario y no puede alterarse por la "experiencia". Creemos que esta ausencia de facultades de adaptación constituye un serio inconveniente. La mejor estrategia general no consiste, pues, en programar directamente un ordenador con la abundancia de pormenores descriptivos de que consta una lengua natural, sino en introducirle el conjunto básico de expectativas y facultades necesarias para aprender una lengua.

Es difícil predecir hasta qué punto conseguirán al fin estas estrategias de investigación aproximarse a la comunicación del habla natural. Un cierto éxito está garantizado, pero no hay duda que habrá que aplicar mucha sabiduría.

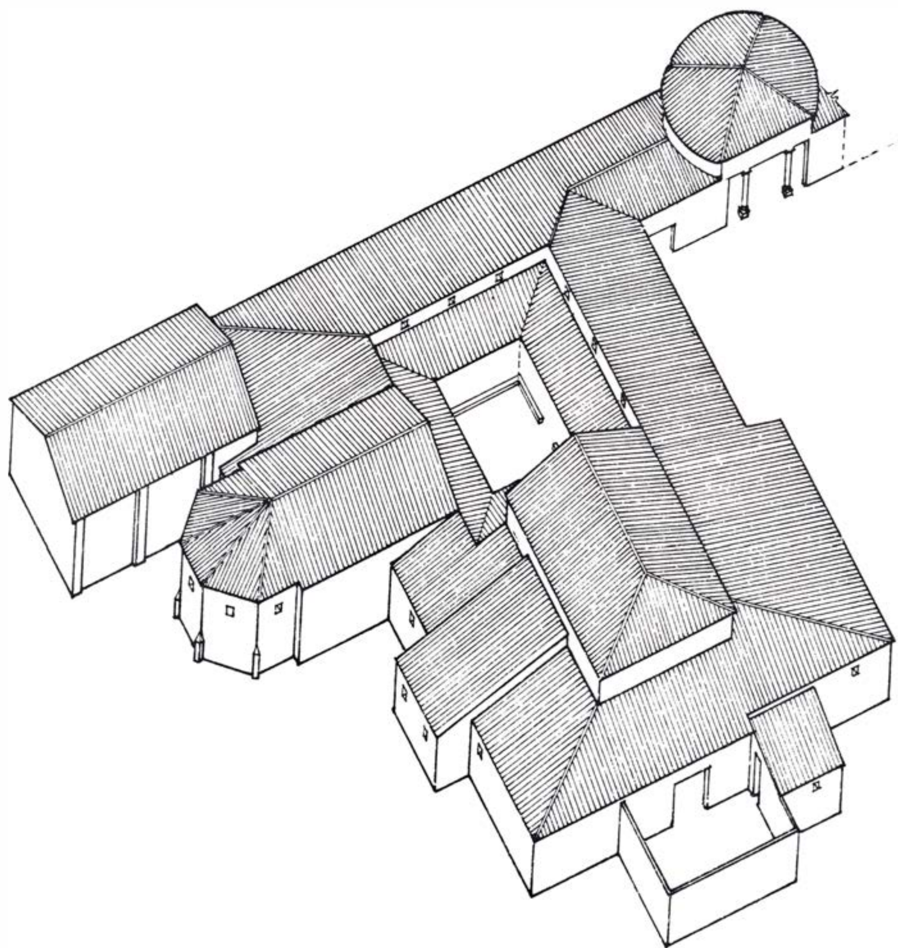
Ciencia y sociedad

La villa rústica romana

En los mosaicos romanos de Toledo, de Centcelles (Tarragona), de Arróniz (Navarra), se divisan en lontananza unos atrayentes edificios campestres de porte señorial, con sus torres, sus galerías o miraderos, sus sombreados pórticos. Son las villas, las casas de campo de los romanos. Ellas cambiaron la fisonomía de Occidente tanto o más que lo hicieron las ciudades romanas, por su mayor dispersión y su tenaz persistencia. La reconstrucción ideal de algunas de las villas (63 en total) reunidas por la doctora Cruz F. Castro en su tesis en curso de impresión traen a las mientes del observador casas monásticas que uno cree recordar. Y nada tendría de extraño, en efecto, que algunas de las primitivas formas de monacato –ortodoxas o heterodoxas, como la de los priscilianistas– se cobijasen en villas como éstas.

La villa es una institución propia del capitalismo agrario de los romanos en todo el Occidente, desde el Rhin al Sáhara, desde el Mar Negro al Atlántico. Unas villas pueden diferenciarse de otras por su tamaño y por su forma; las hay pequeñas como la granja de un modesto labriego y las hay grandes como pueblos (muchos pueblos de Andalucía y de otras regiones –Constantina, Valencina, etc.– conservan los nombres de poseedores de villas romanas); pero todas ellas tienen en común su función agrícola, y no la de casas de recreo o de reposo vacacional. No importan los lujos suntuarios que se permitan; el destino fundamental de la villa será siempre el de una casa de labor.

El primer requisito para su implantación será el de que la tierra sea buena. "Cuando vayas a comprar una finca –aconseja Catón a los lectores de su manual *de agricultura*– visita varias veces el lugar elegido, y mira bien a tu alrededor... Asegúrate de que tiene buen clima, no propenso a tormentas. Que el terreno sea bueno, con fortaleza natural. Si fuese posible, debería hallarse al pie de una colina, orientado a mediodía, en un lugar sano y donde resulte fácil encontrar peones. Debe tener agua abundante y hallarse cerca de una población floreciente, o del mar o de un río navegable, o de una calzada buena y frecuentada." Obsérvese cómo la existencia misma de la villa depende de la ciudad; sin la demanda de alimentos que ésta hace, la villa no podría subsistir, de modo que en caso de que la ciudad no esté próxima, habrá de disponer de otros cauces para dar fácil salida a sus productos.



Reconstrucción de volúmenes de la villa romana de Almenara de Adaja (Valladolid), según Cruz Fernández Castro

Pero sigamos con Catón, que escribe a comienzos del siglo II a.C. cuando el latifundismo romano extendía sus tentáculos por toda Italia: "Si me preguntas cuál es la finca ideal, te diré que la de cien yugadas (250.000 metros cuadrados) de extensión y dotada de toda clase de suelos. Lo primero ha de ser la viña, si produce vino de buena calidad; lo segundo, un huerto irrigado; lo tercero, un saucedal; lo cuarto, un olivar; lo quinto, un prado; lo sexto, un campo de trigo; lo séptimo, un bosque; lo octavo, una arboleda; lo noveno, un encinar" (*Op. cit.* I, 1, 3 y I, 7, 1).

Pese a la variedad de aspectos de su producción, la villa, que está concebida fundamentalmente como un negocio, mira sobre todo a los tres productos de la llamada "tríada mediterránea": el vino, el aceite y el trigo (los dones de Baco, Minerva y Ceres). La producción de vino acumuló tales excedentes a finales del siglo I d.C. (mucho después de Catón, por tanto) que Domiciano decretó la tala de la mitad de los viñedos de las provincias productoras, medida de reconversión cuyos beneficios no tardaron en hacerse sentir. Pero al tiempo que se orientaba hacia estos cultivos, la villa debía ser autosuficiente, para alimentar a los esclavos y al resto de su personal, amén de proporcionar al dueño una renta proporcional a sus inversiones. El dueño (*dominus*) y su familia residían habitualmente en la ciudad y sólo por temporadas en la villa. Su agente permanente en ésta era el capataz, el *villicus*.

Dada la especialización de la villa en los cultivos antes indicados, no tiene nada de extraño que sólo algunas regiones de la Península Ibérica atrajesen en un principio a los inversores romanos, especialmente los valles del Guadalquivir y del Ebro. A comienzos de nuestra era, Estrabón incorpora a su Geografía el relato de un testigo ocular que remonta en una nave el curso del Guadalquivir hasta la ciudad de Córdoba. Las

riberas del río se hallan densamente pobladas. Los campos y los islotes que el viajero encuentra aquí y allá están cultivados con esmero. Los bosques y otras plantaciones contribuyen a la amenidad del paisaje. A mano derecha del que sube, se extiende una llanura dilatada y alta, fértil, en la que se suceden las arboledas y los excelentes pastos. Es el campo de un país civilizado, como debe ser. Unos párrafos más adelante, el geógrafo enumera las principales exportaciones de estas fincas modélicas: en primer lugar, el trigo, el vino y el aceite, "que la Bética no sólo produce en cantidad, sino de la mejor clase", dice textualmente. Aquí radica el por qué de la precoz romanización de Andalucía: su aptitud para la producción agrícola más cotizada por el capitalismo romano. El proceso estaba tan adelantado, a pesar de lo temprano de su fecha, que, según el mismo autor, los andaluces hablaban ya correctamente el latín y habían olvidado sus lenguas vernáculas: "Y poco falta –recalca– para que sean ya romanos todos".

Los romanos habían aportado a Hispania unas ciudades de fisonomía nueva: no encaramadas en alturas como la mayoría de las poblaciones del país, sino asentadas en llanos o en suaves lomas, trazadas con la misma regularidad y provistas de las mismas defensas que un campamento militar (el terraplén y el foso de la Itálica primitiva los hemos detectado en el olivar de Los Palacios, a unos tres metros de profundidad). Aún hoy da gusto andar por ciudades que, como la vieja Zaragoza, conservan el trazado romano de sus calles: rectas, tiradas a cordel y perpendiculares unas a otras, dando lugar a manzanas rectangulares. Pero los habitantes de estas ciudades, representantes de la potencia dominadora, eran a la par terratenientes –"los campesinos, decía Catón, son los mejores de los hombres, ellos dan los mejores ciudadanos, los buenos políticos, los mejores soldados"–, y en su calidad de tales proyectaban sobre el campo concepciones radicalmente nuevas, tanto o más que las ciudades mismas: no sólo villas, sino antes que éstas, calzadas, puentes, estaciones de postas, etc., toda la infraestructura de una campiña civilizada que las villas necesitaban como requisito previo a su existencia misma. Así comenzaron a cambiar al unísono los campos y las ciudades hispánicas; así caminaba con paso firme la romanización del país.

Como es bien sabido por sus graves repercusiones sociales, la tendencia de los latifundios a absorber a los minifundios acarreó en Italia la desaparición del pequeño labrador y el incremento desastroso del proletariado urbano. Es probable que en ciertas regiones peninsulares ocurriese lo mismo. El historiador Plutarco recuerda a un cierto Vibius Paciaecus (un antecesor del apellido Pacheco) que acogió a Craso, huido de Roma, en una propiedad de la zona de Málaga que por su descripción debía de ser inmensa, con parajes que sólo algún pastor visitaba muy de tarde en tarde. Y esto a comienzos del siglo I a. C., o sea, muy pronto. El efecto más palpable de este latifundismo sería el mismo que se percibe hoy en Carmona, Lebrija y otros puntos de Andalucía: que los habitantes de muchos núcleos de población dependerían de las villas para su subsistencia o tendrían que emigrar a ciudades grandes como Sevilla para trabajar allí como obreros portuarios, barqueros, bomberos, etc. (estos eran los *scapharii, lyntrarii, centonarii,* etc., de Cánama, de Oducia, de Naeva y de otras poblaciones ribereñas que figuran en las inscripciones romanas de Sevilla).

Si buena parte de los latifundistas de Andalucía, Cataluña y Aragón residían la mayor parte del año en algunas de las muchas colonias y municipios romanos existentes en sus territorios, es probable que en regiones como Castilla, Navarra y León, donde las ciudades eran menos y más alejadas unas de otras, su residencia permanente se encontrase en la villa rústica. Esto explicaría las dimensiones y la suntuosidad de las llamadas villas señoriales que aquí florecen en el Bajo Imperio. La conversión de la villa señorial o dominical en un centro del que dependían otras villas subsidiarias imponía servicios tales como baños multitudinarios, graneros de alta capacidad, enormes hornos, etc. Aquí se ha querido ver el germen del colonado de la Alta Edad Media, que permitiría a un sujeto hispanorromano como el suegro del visigodo Teudis (comienzos del siglo VI) reclutar, en un momento dado, 2000 lanceros entre sus colonos. (Antonio Blanco Freijeiro.)

Astrología hispánica hacia el año 800

En 1961 dos hispanistas norteamericanos, Lloyd A. Kasten y Lawrence B. Kiddle, publicaron una edición crítica del *Libro de las Cruzes* de Alfonso X el Sabio que ponía al alcance de los eruditos del mundo entero un texto alfonsí aún inédito, puesto que no formaba parte de los célebres *Libros del Saber de Astronomía.* A pesar de que esta obra era de fácil acceso, no ha despertado hasta muy recientemente el interés de los historiadores de la astronomía. La actitud predominante entre éstos ha sido resumida, de manera lapidaria, por el gran maestro Otto Neugebauer, en su monumental *History of Ancient Mathematical Astronomy* (1975), el cual afirma que el libro contiene "una enumeración interminable de combinaciones triviales de influencias astrológicas lo que revela [por parte de su autor] una torpeza de mente poco usual".

La lectura del texto castellano ofrecía, no obstante, algunos datos interesantes. En el prólogo se afirmaba que el libro, en versión árabe, fue "hallado" por el rey Don Alfonso quien lo hizo traducir al castellano por Yehudá b. Moshé (fl. 1225-1276) con la colaboración de Johan Daspa y que, a su vez, la versión árabe era una reelaboración de un texto "antiguo" realizada por un tal "Oueydalla" (Ubayd Allah), identificado conjeturalmente por Millás Vallicrosa con Abu Marwán Ubayd Allah b. Jalaf al-Istidjí, astrólogo del siglo XI. El *Libro de las Cruzes* suele referirse a "Oueydalla" como "el esplanador" y hace hincapié en el hecho de que este autor "halló" el libro, lo reescribió y lo explicó dejándolo en su forma actual. Existía, pues, una versión anterior del mismo a la que alude el segundo prólogo del libro, escrito por el propio Oueydalla: en él se pone de manifiesto que la versión antigua correspondería a una tradición astrológica occidental –"africana" e hispano-romana– distinta y menos elaborada que la oriental (babilónica, griega, persa y árabe).

Así estaban las cosas cuando Juan Vernet, en 1971, y Rafael Muñoz, en 1979, dieron a conocer dos manuscritos árabes conservados en El Escorial, que contenían pasajes de la versión árabe del *Libro de las Cruzes.* El manuscrito descubierto por Vernet tenía una importancia especial no sólo porque confirmaba plenamente las afirmaciones básicas del prólogo de Oueydalla, al que acabo de aludir, acerca del carácter occidental de la tradición astrológica representada por el *Libro de las Cruzes,* sino sobre todo porque los pasajes árabes del mismo terminan con la cita de treinta y nueve versos de un poema didáctico escrito por Abd al-Wahid b. Ishaq al-Dabbí que constituyen una versificación del capítulo 57 del *Libro.* Ahora bien, sabemos que este al-Dabbí fue astrólogo de corte del emir cordobés Hisham I (788-796) y, por consiguiente, la versión más antigua conoci-

da del *Libro de las Cruzes* remonta, por lo menos, a finales del siglo VIII o principios del siglo IX. Nos encontramos, pues, ante el hecho importante de que la obra alfonsí es una reelaboración del texto astrológico hispánico más antiguo conocido. Por otra parte Vernet, en su trabajo de 1971, ya señaló que a principios del siglo IX no se había producido aún la introducción en la España Musulmana de ningún texto astrológico árabe oriental de tradición helenística. Este detalle unido a la insistencia, tanto del texto alfonsí como del texto árabe conservado, en que el sistema de las cruces era el antiguo sistema astrológico utilizado en España y en el Norte de Africa y en el que no se utilizaban las sutilezas orientales, llevó a Vernet a una conclusión inevitable: el sistema de las cruces parece de origen latino y, anterior a la versión de al-Dabbí, debió existir un texto astrológico bajolatino conocido en la España Visigoda. Esta conclusión es muy defendible si recogemos las alusiones a la difusión de la astrología en nuestro país en tiempos de Isidoro de Sevilla (c. 560-636): a pesar de la lucha oficial que el obispo sevillano mantuvo contra las convicciones astrológicas, es obvio que en su obra quedan restos de este tipo de creencias que pueden atribuirse a la persistencia de la herejía priscilianista –que mantenía dogmas astrológicos–, a la presencia en la Bética de fenicios, cartagineses y sirios que seguían practicando religiones astrales, a la permanencia de judíos helenizados que habían conciliado su fe con la astrología y, finalmente, a influencias bizantinas.

Podría, aquí, continuar describiendo las técnicas de predicción astrológica utilizadas en el sistema de las cruces y mostrar en qué sentido son más rudimentarias que las que se encuentran en la tradición greco-árabe. Prefiero, no obstante, mostrar cómo el estudio de un texto astrológico muy primitivo como el *Libro de las Cruzes* puede proporcionar información acerca de los procedimientos de cálculo astronómico utilizados por los astrólogos. Dicho de otro modo: pretendo hacer ver, a través de un ejemplo concreto, cómo la astrología tiene un evidente interés para la historia de la astronomía, ya que suministra datos sobre épocas particularmente obscuras. El problema a considerar es muy simple: sabemos que a fines del siglo VIII existen en España astrólogos en ejercicio que levantan horóscopos característicos según el sistema de las cruces. Ahora bien, para levantar un horóscopo es preciso calcular las posiciones de los "planetas" conocidos (Sol, Luna, Mercurio, Venus, Marte, Júpiter y Saturno) en un momento determinado. A lo largo de la Edad Media estas posiciones se calcularon utilizando tablas astronómicas, efemérides anuales, almanaques perpetuos o computadores analógicos denominados ecuatorios. Ninguno de estos medios eran accesibles a un astrólogo hispano a principios del siglo IX: hacia el 850 se introducen en la España Musulmana las primeras tablas astronómicas; las efemérides anuales eran conocidas en Oriente desde el siglo X, pero no hay evidencia alguna de que fueran nunca utilizadas en la España Medieval; los almanaques perpetuos y los ecuatorios, en cambio, son desarrollos que parecen de origen hispánico, pero surgen en el siglo XI. El problema, por tanto, sigue en pie y puede reducirse a la pregunta siguiente: ¿podían nuestros primitivos astrólogos levantar horóscopos contando con los dedos?

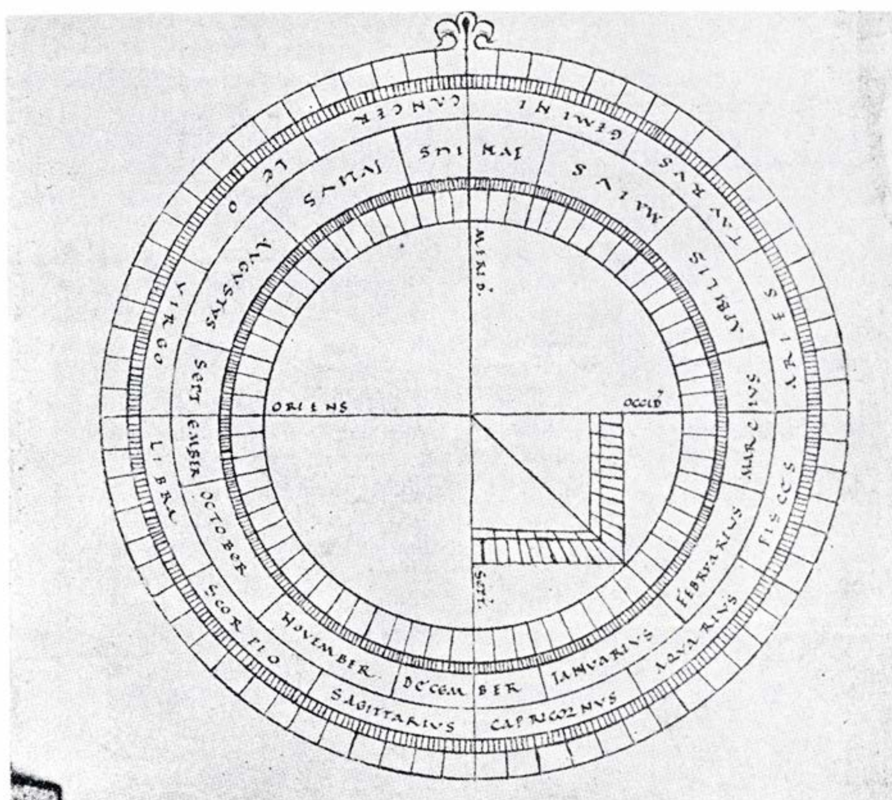
Creo que podemos encontrar una respuesta a esta cuestión si recurrimos a la rudimentaria tradición astronómica latina representada por el *Libro de las Cruzes* y por la literatura de cómputo eclesiástico. En primer lugar, los horóscopos levantados de acuerdo con el sistema de las cruces no exigen precisión alguna: basta con determinar el *signo zodiacal* en el que se encuentra el planeta, lo que implica una tolerancia de error que puede llegar a 30 grados. Por otra parte este sistema insiste sobre todo en la relevancia de la posición de los dos planetas más lentos: Saturno (cuyo período sidéreo es de unos treinta años) y Júpiter (con un período sidéreo de unos doce años). En tercer lugar, en un curiosísimo pasaje del *Libro de las Cruzes,* Oueydalla hace una serie de reproches a los astrólogos que levantaban horóscopos basándose únicamente en las posiciones medias –no las verdaderas– de los planetas. Si aceptamos que estos astrólogos que recurrían sólo a las longitudes medias para hacer sus predicciones son probablemente los partidarios del antiguo sistema de las cruces que Oueydalla trataba de corregir y hacer más preciso introduciendo en él ciertos refinamientos propios de la astrología oriental, tendremos una pista para aclarar los métodos de cálculo que utilizaban. En efecto, los tratados de cómputo eclesiástico contienen reglas muy elementales y diagramas que permiten calcular aproximadamente la longitud media del Sol y de la Luna y,

Tabla cuadrática de los signos zodiacales en un manuscrito del siglo x procedente de la abadía de Ripoll

en ocasiones, las de los restantes planetas. Un escrito atribuido a Beda titulado *De planetarum et signorum ratione* contiene una doble serie de reglas de esta índole aplicables a los cinco planetas propiamente dichos. No obstante, como ejemplo de esta hipótesis, traeré a colación únicamente el caso del Sol y de la Luna, ya que son los dos astros que más interesan a los tratadistas de cómputo, preocupados por los problemas que plantea el calendario lunisolar eclesiástico.

En lo que respecta al Sol, la astronomía árabe conoció reglas para determinar, aproximadamente, su longitud media. Métodos de esta índole resultan tan sencillos en el caso del Sol que es más que probable que hubieran sido conocidos por el cómputo visigótico, aunque no puedo demostrarlo. En cambio resulta obvio que esta tradición latina sí conoció tablas o diagramas que establecían una correspondencia biunívoca entre la fecha del año juliano y la longitud del Sol y ello está documentado en manuscritos latinos hispánicos por lo menos desde el siglo IX. Un indicio de que este tipo de procedimientos para determinar, aproximadamente, la longitud media del Sol pudo ser transmitido al mundo hispanoárabe lo tenemos en los calendarios zodiacales que aparecen, de manera característica, en los astrolabios hispano-árabes y norteafricanos. Este diagrama es muy elemental: estos instrumentos suelen llevar, en su reverso, dos círculos no concéntricos en uno de los cuales –dividido en 365 partes– se encuentran representados los doce meses y los días del año, mientras que en el otro –dividido en 360 grados– aparecen los doce signos zodiacales, cada uno de los cuales consta de 30 grados. Resulta fácil de comprender que, si ambos círculos están correctamente trazados, podemos utilizar la alidada del astrolabio como regleta y establecer, con enorme facilidad, la longitud del Sol para cada día del año. Este dispositivo está documentado en España desde el siglo X y su origen fue discutido, hace bastantes años, por Zinner (1944) y por Millás (1947). A la vista del conjunto de datos que conocemos hoy, cabe plantearse si, realmente, puede tener raíces latino-mozárabes y entroncarse con los tratados de cómputo antes aludidos.

Para determinar la longitud media de la Luna disponemos de una regla simple conservada en un manuscrito del Museo Diocesano de Vich fechado en 1235. En ella se establece lo siguiente: se toma la longitud de la Luna en el momento de su conjunción con el Sol



Dorso de un astrolabio latino en el que aparece un calendario zodiacal

(Luna nueva), en el que la posición de ambos astros será forzosamente la misma. A continuación se obtiene la "edad de la Luna" (o sea el día del mes lunar), cifra que se multiplica por 4 y se divide por 10. El resultado se suma a la posición de la Luna en la última Luna nueva y se obtendrá el signo zodiacal en que se encuentra nuestro satélite en el día de referencia.

En efecto, el mismo manuscrito nos indica que la Luna recorre diariamente 13;10,35° ya que su mes trópico es de 27 días, 7 horas y 45 minutos. Ahora bien, el parámetro 13;10,35° expresado en signos zodiacales equivale a:

$$\frac{13;10,35^{\circ}}{30^{\circ}} = 0{,}44 \text{ s.}$$

(en notación decimal). Por tanto resulta aceptable el que la regla del manuscrito de Vich establezca que hay que multiplicar la edad de la Luna por 4 y dividirla por 10, ya que esto equivale a multiplicar por 0,4, obteniéndose el resultado en signos zodiacales.

Ahora bien, la regla anterior sólo aparece documentada en fuentes tardías. No sucede lo mismo con la llamada "tabla cuadrática de los signos zodiacales" que remonta, por lo menos, al siglo VIII. En España está documentada, por lo menos, desde el siglo X y debió ser muy popular en la Baja Edad Media, ya que aparece con frecuencia en textos hispánicos de los siglos XIV y XV. La característica externa más notable de esta tabla es que, en ella, los signos zodiacales aparecen repetidos en diagonal. Su uso es simple y se basa en dos principios erróneos: 1) La Luna recorre un signo zodiacal en 2,5 días, lo que equivale a afirmar que nuestro satélite recorre diariamente sólo 12°, parámetro que, según parece, deriva del cómputo eclesiástico; 2) El principio de cada mes coincide con la entrada del Sol en un signo zodiacal (el 1.° de marzo en Aries, el 1.° de abril en Tauro, etcétera). Sobre las dos bases anteriormente citadas y teniendo en cuenta que, en la tabla, la hilera inferior horizontal lleva los nombres de los doce meses del año y la columna de la izquierda tiene distribuidos, en doce casillas, los 29 o 30 días del mes lunar, se empieza por averiguar cuántos días han transcurrido desde la Luna nueva y se busca el número correspondiente en la columna de la izquierda. Se considerará, luego, el mes del año en la hilera inferior horizontal: la intersección de ambas hileras, horizontal y vertical, indicará la casilla correspondiente al signo zodiacal de la Luna. Obviamente, el resultado será sólo vagamente aproximado, pero el procedimiento podía ser

utilizado con propósitos meramente astrológicos.

Lo anterior puede servir como ejemplo de la existencia, en la tradición computística latina altomedieval, de métodos aproximativos para calcular la longitud media de los astros. Estos métodos pudieron ser conocidos por los astrólogos hispano-árabes en una época en la que aún no se había producido la recepción de la astronomía oriental, pero no debieron persistir mucho más allá del 850. No obstante, la tradición astrológica representada por el *Libro de las Cruzes* se mantuvo viva hasta el siglo XI y, posiblemente, hasta el XIII. (Julio Samsó.)

Investigación y desarrollo en hortofruticultura

En las dos colaboraciones anteriores (INVESTIGACIÓN Y CIENCIA, abril y mayo de 1981) nos hemos ocupado de la producción hortofrutícola de exportación y de los principales problemas que tienen planteados los cultivos de mayor interés económico. En ésta, consideraremos las acciones que deberían emprenderse para contribuir a la resolución de los problemas enumerados. En España disponemos de medios materiales y humanos suficientes para abordarlos y con muchas posibilidades de llegar a resolverlos satisfactoriamente. Señalaremos las acciones a emprender para mejorar la competitividad de la hortofruticultura de exportación.

Comencemos, una vez más, por la citricultura. Para paliar los problemas reseñados sería conveniente proceder a la selección de estirpes de naranjo amargo tolerantes a la tristeza de los cítricos, así como a la introducción y adaptación de otros portainjertos, la potenciación de los programas (actualmente en marcha) para la preparación de material vegetal de las diferentes especies y variedades de agrios libres de virus, desarrollo de métodos eficaces de preinmunización de cítricos contra la tristeza, caracterización de los agentes causantes de otras virosis (psoriasis, xiloporosis, impietratura, etc.) de los cítricos, estudio de los mecanismos de replicación y expresión de síntomas del viroide causante de la exocortis y, finalmente, aislamiento y caracterización de las micoplasmosis, particularmente, la causante del Stubborn de los cítricos.

Una buena aportación al mejor uso de los fertilizantes y del agua de riego podría conseguirse mediante el estudio de la absorción, traslocación, acumulación y movilización de las reservas de nitrógeno y fósforo en los cítricos a lo largo del ciclo vegetativo, acompañado del estudio del comportamiento de los fertilizantes en el suelo (pérdidas de N, P y K por lixiviación, volatilización o fijación irreversible en el suelo) y de la determinación de las necesidades de agua de las diferentes especies y variedades de cítricos, así como de la eficacia de los diferentes sistemas de riego. En lo que respecta a la fisiología, conviene destacar la necesidad de investigar: la influencia del equilibrio hormonal sobre la floración y cuajado de los frutos (mecanismo de acción hormonal en la fructificación), el proceso de abscisión de frutos (acción de las fitohormonas y enzimas implicados en el mismo) y el de maduración de los cítricos (mecanismo de acción del etileno).

No debe olvidarse la necesidad de proceder a la caracterización de espe-

cies de insectos parásitos de la mosca blanca (*Aleurothrixus floccosus*) y de la mosca del mediterráneo (*Ceratitis capitata wied*), la reproducción y estudio de las posibilidades de su empleo en programas de lucha biológica y el conocimiento de los microorganismos patógenos que afectan a los frutos antes y después de la recolección. Importa no descuidar todas aquellas acciones relativas a la conservación post-recolección (atmósfera controlada, refrigeración) de los frutos cítricos.

Para contribuir a la resolución de los problemas que tienen planteados los frutales distintos de los cítricos creemos que debieron acometerse las siguientes acciones: *1.ª*) En el caso de los frutales de hueso, se impone proceder al aislamiento e identificación de los micoplasmas causantes de las alteraciones "peach rosete" y "apricot clorotic leaf roll", y desarrollar formas de lucha que permitan combatirlos y prolongar así la vida productiva del arbolado. *2.ª*) En los productores de frutos secos, se han de encontrar tratamientos fungicidas económicos para combatir las exudaciones gomosas del almendro. Los problemas del avellano, de mayor complejidad, deben abordarse a partir de selecciones de patrones adecuados a su ecología y desarrollo de técnicas de cultivo que mejoren la precocidad de la producción. *3.ª*) Uno de los problemas a resolver en el cultivo de frutas tropicales (aguacate y chirimoyo) es la introducción de variedades y portainjertos aptos para las condiciones ecológicas de la costa del Mediterráneo y de las islas Canarias. Hay que definir el área de cultivo de las especies más importantes. En el caso del aguacate, se aconseja introducir variedades que puedan garantizar una producción en los meses de verano para completar el ciclo anual de producción. En la escasa superficie que se está dedicando a estos cultivos en España se ha detectado un problema importante causado por hongos del género *Phytophthora* (que pudre las raíces), cuya resolución no puede postergarse. La fructificación del chirimoyo no se produce con la regularidad deseada para garantizar la rentabilidad del cultivo. La corrección de esta irregularidad deberá abordarse estudiando la fisiología del proceso de fructificación, mediante la aplicación de hormonas, o el establecimiento de polinizadores adecuados. Una vez más, conocer la fisiología de la floración y de la fructificación es condición necesaria para establecer las técnicas de cultivo que mejoren su productividad. *4.ª*) Somos de la opinión de que habría que profundizar en los esfuerzos para introducir nuevos frutales de origen tropical, como son el litchi, ananás, macademia, mamey y flor de pasión, así como dominar los frutales de cultivo tradicional que no son objeto de comercio exterior: mango, papaya y guayaba.

En lo referente a la horticultura, los autores resumen así las medidas que deberían tomarse. En primer lugar, la estructuración de los equipos de investigación en el campo de la mejora vegetal, coordinándolos a nivel nacional, para obtener variedades de hortalizas que satisfagan las exigencias de calidad del mercado europeo y al mismo tiempo que resistan las plagas y enfermedades de mayor impacto económico. El establecimiento de la estructura investigadora que evitara la importación de semilla de patata justificaría por sí sola

la necesidad de la misma. Una aproximación al problema del "cansancio del suelo" pasa por la realización de estudios muy diversos, tales como los de poblaciones de nemátodos y microorganismos del suelo y fitotoxinas que dejan los diferentes cultivos, cuya caracterización exige un trabajo notable dentro del campo de la química orgánica y la fisiología vegetal. Por lo que concierne a las pérdidas de productos agrícolas post-recolección, habría que replantearse la atención prestada al conocimiento de los patógenos y a las formas de lucha contra los mismos. Ahondando, además, en las tecnologías auxiliares, tales como refrigeración y atmósfera controlada, y mejorando la infraestructura necesaria para controlar y garantizar la salubridad de los productos agrarios (definir calidades, determinar residuos de plaguicidas, etc.).

Y llegamos así al capítulo de las plantas ornamentales. En este sector, y cubriendo un abanico de soluciones que va desde la selección de variedades y técnicas de cultivo más adecuadas hasta la fisiología post-recolección y transporte a los mercados consumidores, las acciones a emprender se podrían sintetizar como sigue: *a*) Estudio de las condiciones más adecuadas para el cultivo de especies autóctonas que permitan obtener productos competitivos. Lo que comporta un aumento de la gama de productos comercializables a fin de diversificar la producción. *b*) Mejora de las condiciones sanitarias de las plantas cultivadas, por conocimiento de las plagas y enfermedades de las diferentes especies, así como de los tipos de tratamientos más adecuados para combatirlas. El estado sanitario de los bulbos (tulipanes, jacintos, gladiolos, etc.), debido a su forma de reproducción, es todavía muy deficiente. *c*) Mejora vegetal de plantas ornamentales. Al igual que en horticultura, conviene establecer la infraestructura necesaria para proceder a una mejora de las especies y variedades de ornamentales cultivables, tanto para modular la gama de productos comercializables como para disponer de variedades resistentes a plagas. *d*) Estudio de conservación post-recolección de las ornamentales. Para garantizar el mantenimiento de la calidad, hay que conocer la fisiología del producto y las plagas que pueden afectarle. *e*) Estudio de las causas que provocan malformaciones de las flores, clavel, rosa, etc. Especial atención merece el estudio de los ciegos "blind shoot" en el rosal y cuya solución podría suponer, en muchas explotaciones, mejoras de hasta un 50 por ciento.

La relación de problemas, así como la propuesta de acciones a emprender para resolverlos, da una idea de la magnitud de la tarea que resta. Por ser ingente, parece conveniente proceder a la definición de objetivos prioritarios, que variará de acuerdo con el criterio seguido, no siendo el menor el de índole económica. Es evidente que los programas de investigación que se aborden deben adaptarse a las posibilidades reales de España y constituir un factor eficaz para su inmediato desarrollo. En esa prelación de criterios, pondríamos en primer puesto los conocimientos básicos en que ha de apoyarse el estudio a emprender. Puede ser engañoso pensar que el retraso científico radica en que los investigadores españoles no participamos en los últimos descubrimientos o desarrollo científicos, pues el retraso no esta ahí, sino en que no se dominan y aplican técnicas existentes y conocidas desde hace muchos años. Señalaríamos luego la disponibilidad de técnicas y personal necesario convenientemente entrenado: en ocasiones, algunos temas de gran interés científico y repercusión económica-social deberán rechazarse por no disponerse del mínimo de personal preparado, lo que no exime de la necesidad de formarlo.

Destacaríamos en tercer lugar la proyección de los resultados en la sociedad a la que debemos servir. Uno de los frenos del desarrollo de la investigación agraria de España ha sido la falta de proyección de los resultados de la investigación en la sociedad. Unas veces por la índole de los avances logrados que tienen interés puramente académico y, otras, por falta de la estructura necesaria para trasladar los resultados del laboratorio al campo, los progresos realizados no han trascendido a la sociedad que, siendo la financiadora de la investigación, quiere conocer el fruto de la inversión.

Para estructurar los grupos de investigación que deben abordar los estudios que se definan como prioritarios es necesario realizar un inventario de los recursos humanos y materiales con los que cuenta cada uno de los entes implicados: universidad, ministerios, centros especializados, etcétera. Este inventario debe contemplar, además de las especialidades cultivadas, las técnicas y conocimientos con los que se cuenta. Estos equipos deberían tener suficiente versatilidad y agilidad para poder asimilar los conocimientos alcanzados en otros países y al mismo tiempo imaginación necesaria para avanzar con seguridad en la resolución de los problemas que afectan al país. (P. Cuñat, A. Aguilar y V. García.)

Origen de la información genética

Se han deducido y comprobado las leyes que gobiernan la selección natural de moléculas prebióticas, lo que permite descubrir la interacción de los primitivos genes del ácido ribonucleico (ARN) con proteínas y el origen de la clave genética

Manfred Eigen, William Gardiner, Peter Schuster y Ruthild Winkler-Oswatitsch

Charles Darwin vio en la diversidad de las especies los principios evolutivos que las originan: variación, competencia y selección. Desde la época de Darwin se ha conseguido un conocimiento de la biología molecular y de la geofísica y la geoquímica de la Tierra prebiótica inimaginable en el siglo XIX. ¿Podemos remontarnos ahora en el curso de la evolución hasta la era anterior a la aparición de los organismos?

Una respuesta inmediata sería que no. El registro fósil prebiótico, por lo que sabemos, desapareció o fue destruido por los seres vivos posteriores. Los fósiles intelectuales que quedan (la clave genética, los mensajes genéticos de los organismos actuales y las rutas metabólicas conocidas) dan una información tan fragmentaria que uno nunca podría describir la evolución prebiótica con tanto detalle como, por ejemplo, la evolución de los primates.

Pero lo fragmentario de la información no ha sido nunca una barrera al descubrimiento de leyes naturales. Newton descubrió las leyes universales del movimiento a partir de observaciones de unos pocos planetas; Mendeleev descubrió la estructura de la tabla periódica en la química de sólo unos pocos elementos; los físicos de hoy infieren leyes que describen las interacciones de partículas elementales a partir de la observación de un reducido número de sucesos. No hace falta conocer con detalle las condiciones y los hechos prebióticos para descubrir las leyes evolutivas que condujeron a la primera vida sobre la Tierra. Sólo hay que confiar en que queden suficientes indicios fósiles para guiar el propio pensamiento y exigir que la teoría tenga suficiente capacidad de predicción como para poder confirmarla experimentalmente. En este sentido, la contestación a la pregunta anterior es afirmativa: se pueden hacer aserciones definidas sobre las leyes naturales que gobernaron el origen y la evolución prebiótica de la vida.

En este artículo presentamos lo que hay que añadir a las ideas de Darwin para describir la evolución anterior a la existencia de organismos. Primero mostraremos que sus ideas valen para la evolución a niveles muy inferiores al de organismo. Para explicar el origen de la complejidad de los organismos superiores y de la diversidad de las especies, Darwin propuso que lo más complejo derivaba de lo menos complejo por selección natural. ¿Por qué no valdría este principio también para la complejidad de las macromoléculas? Enunciaremos las condiciones necesarias y suficientes para que la selección natural actúe a nivel molecular. Esta selección conduce a resultados repetibles que se dan inevitablemente en cuanto se cumplen ciertas condiciones.

La Tierra antes de la vida

Aunque la competencia es la base de la selección natural entre organismos, por sí sola no habría seleccionado, en tiempos prebióticos, las estructuras más aptas; ciertas formas de cooperación también fueron esenciales. La interacción evolutiva de la competencia y la cooperación entre moléculas reflejaba la necesidad de procesar y utilizar la información genética primitiva para estabilizarla y, luego, perfeccionarla. Es imposible repetir los pasos históricos de ese perfeccionamiento porque, durante la evolución primitiva, se probaron y descartaron muchísimas mutaciones aleatorias; pero se pueden entender las leyes naturales que los gobernaron. Estas leyes pueden comprobarse de varias maneras: por experimentos con virus bacterianos, por estudios químicos de los ácidos nucleicos y de las proteínas y por análisis comparativo de ácidos nucleicos y proteínas que han sobrevivido tres o cuatro mil millones de años de evolución.

Antes de la representación del drama de la vida tuvo que organizarse el escenario y aparecer ciertos actores secundarios. El escenario fue algún lugar de la Tierra primitiva, con una temperatura no muy diferente de la actual. La composición de la superficie terrestre también se parecía a la actual, si se considera solamente la abundancia de los elementos, pero era muy diferente en el modo en que se combinaban los elementos. Se ha demostrado experimentalmente que casi cualquier fuente de energía, rayos, ondas de choque, radiación ultravioleta o cenizas volcánicas calientes, habría convertido una parte significativa de los materiales primitivos de la superficie en una gran variedad de sustancias que hoy consideraríamos orgánicas. El sistema solar primitivo comprendía también muchos cometas y meteoritos, que pueden haber contribuido sustancialmente a la conformación de la superficie terrestre. La acción de la radiación solar sobre este material ultrafrío, residuo de la condensación del sistema solar, podría haber producido moléculas orgánicas tan grandes como algunos polímeros biológicos.

Todas las hipótesis sobre la "sopa primordial" de la que surgió la vida coinciden en que no sólo incluía los azúcares, aminoácidos y otras sustancias que son ahora reactivos bioquímicos esenciales, sino muchas otras moléculas que son meras curiosidades de laboratorio. El primer elemento organizativo tuvo que ser, por tanto, muy selectivo desde el principio. Debía tolerar una enorme sobrecarga de pequeñas moléculas biológicamente inútiles pero químicamente posibles. Tenía que seleccionar las moléculas que más tarde serían los monómeros característicos de los polímeros biológicos y unirlas fiablemente en determinadas configuraciones.

La cantidad total posible de materia orgánica era inmensa. Si el carbono

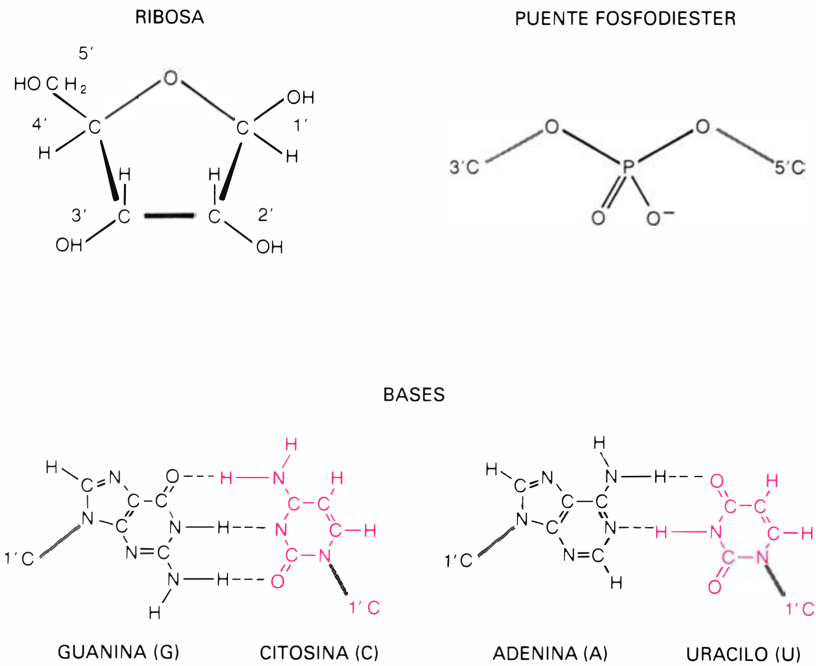
que ahora se encuentra en forma de carbón, carbonatos y materia orgánica estuviera uniformemente distribuido en toda el agua actual de los océanos, resultaría una solución de carbono tan concentrada como una buena sopa. Los procesos geofísicos como la erosión, la evaporación y la sedimentación han de haber actuado, entonces como ahora, creando una diversidad de ambientes. Evidentemente, al menos uno de estos ambientes fue apropiado, en temperatura y composición, para el origen de la vida.

La sopa primitiva se enfrentó a una crisis energética: las formas de vida primitiva tuvieron que extraer de alguna manera energía química de las moléculas de la sopa. Para nuestra historia no importa cómo lo hacían; podemos suponer que existía algún sistema de almacenamiento y liberación de energía basado en fosfatos. La recarga no metabólica de este depósito de energía (quizá por alguna forma de conversión de energía solar en energía química) duraría hasta que apareciera un mecanismo para fermentar algunos componentes de la sopa que no sirvieran para otra cosa. La fermentación debió bastar hasta que la invención de la fotosíntesis proporcionó una fuente continua de energía.

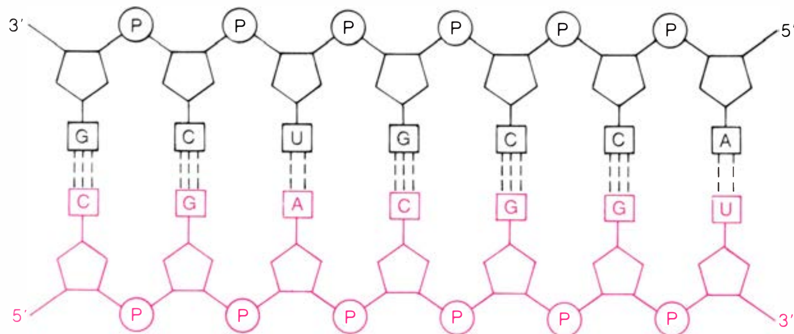
Los primeros genes

En las células, la información genética está contenida en el ADN, que se transcribe en ARN mensajero y se traduce a proteínas; en los virus, la información está en cadenas de ADN o de ARN. Ambos ácidos nucleicos son largas cadenas de nucleótidos. Cada nucleótido tiene tres componentes: un grupo químico llamado base, un azúcar (desoxirribosa en el ADN, ribosa en el ARN) y un grupo fosfato. El esqueleto de la molécula está formado por azúcares y fosfatos unidos; la información genética está contenida en secuencias determinadas de bases. Las cuatro bases del ADN son las purinas adenina

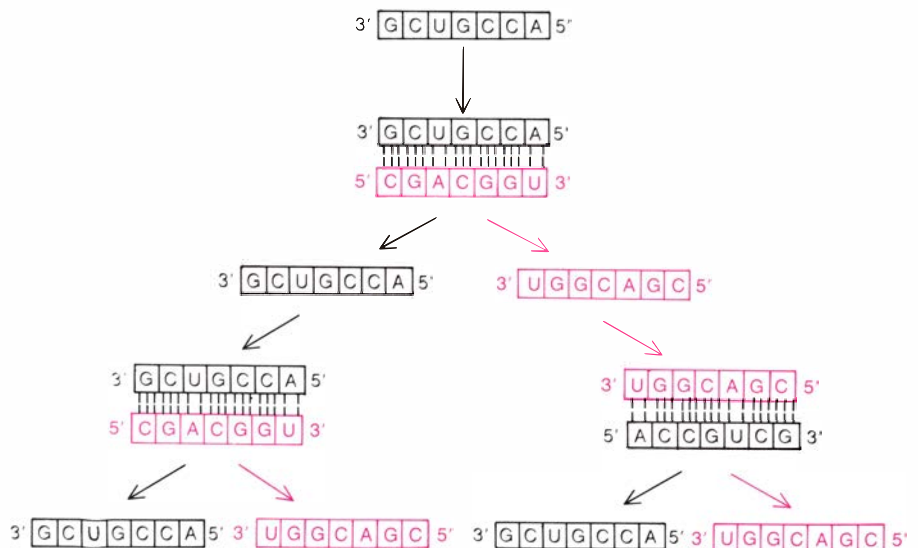
a



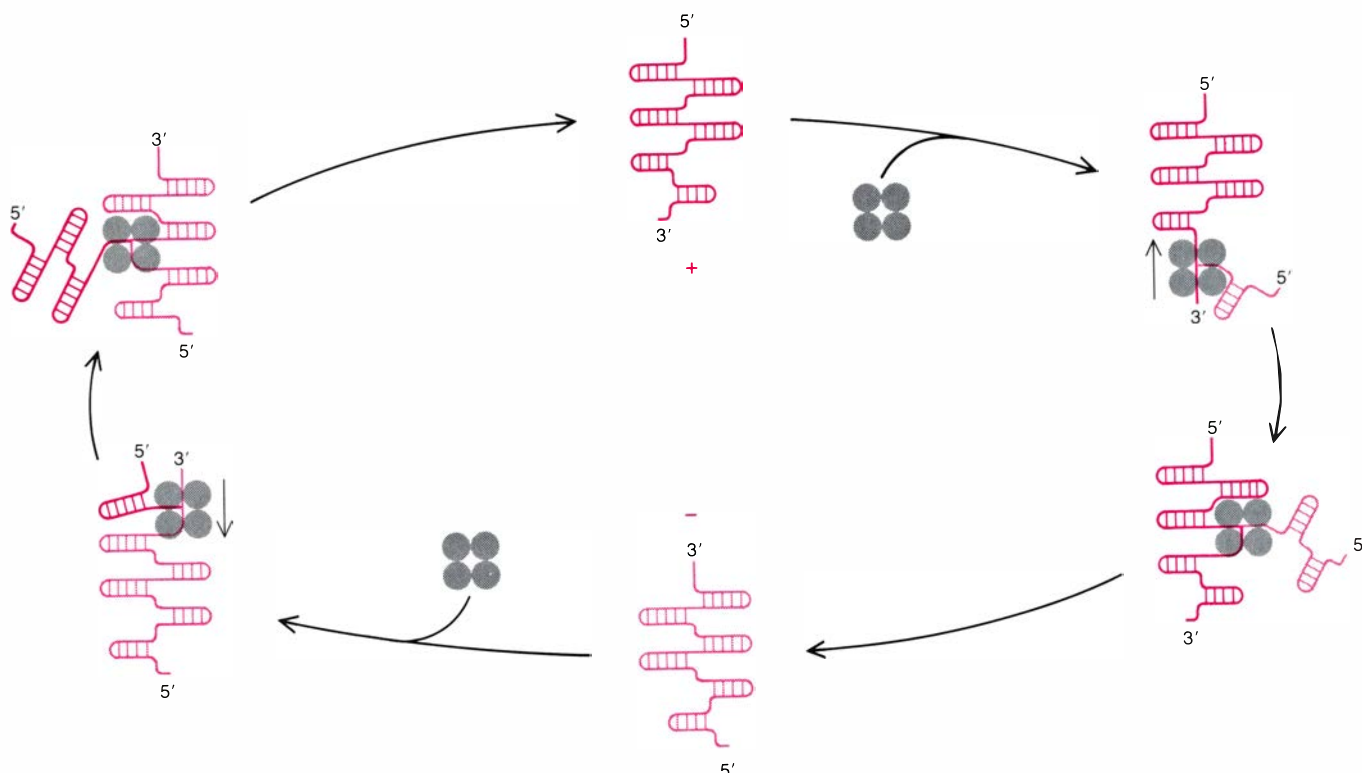
b



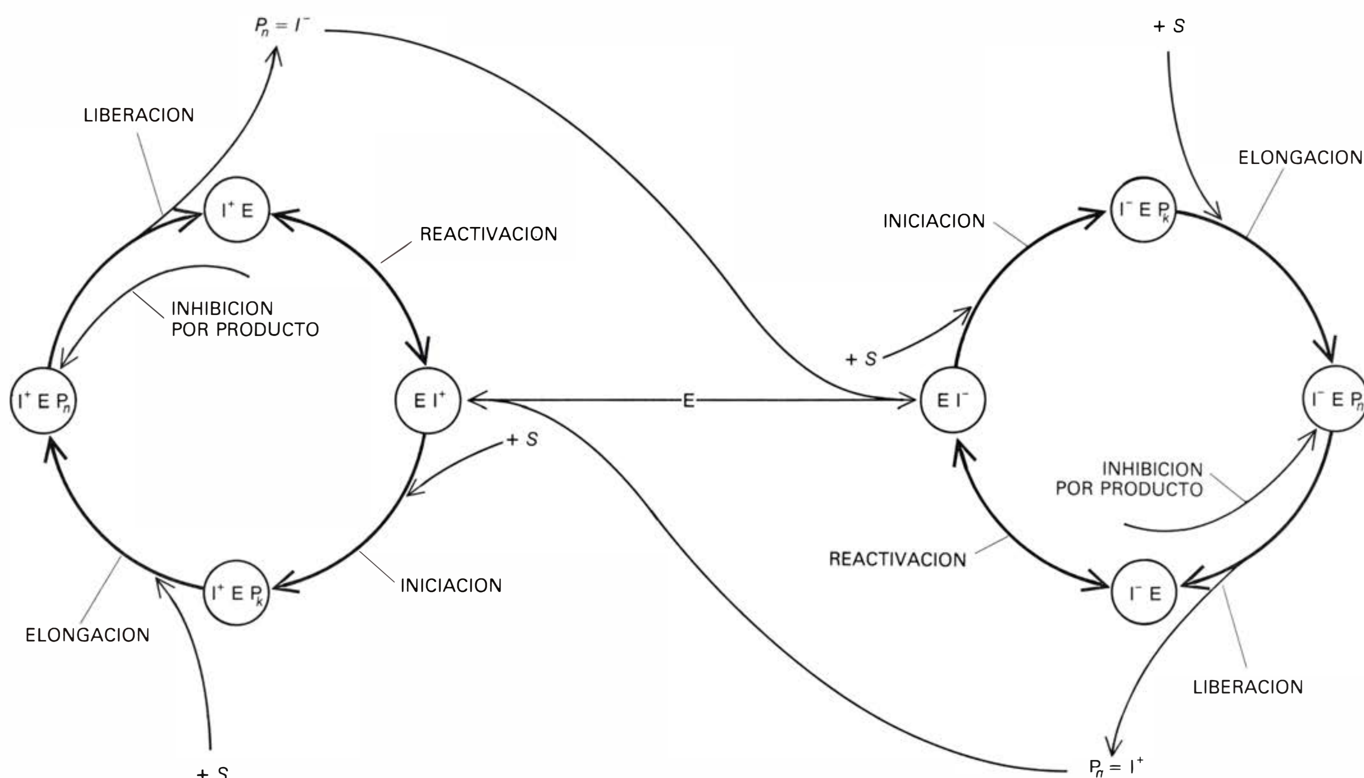
c



PRIMEROS PORTADORES de información genética: eran moléculas de ARN, cadenas lineales de nucleótidos unidos por puentes fosfodiéster. Cada nucleótido del ARN consta de un azúcar ribosa unido a una de las cuatro bases (*a*); la información está codificada en la secuencia de bases a lo largo de la cadena de ARN. Las bases son complementarias: por medio de puentes de hidrógeno (*trazos discontinuos*) la adenina se aparea específicamente con uracilo y la guanina con citosina. La información del ARN se transfiere por síntesis de una cadena complementaria (*color*), para la que la cadena original sirve de molde (*b*). Se muestra también la transferencia (*c*) de información a lo largo de dos generaciones de una secuencia de 7 nucleótidos.



EL ARN UNICATENARIO del virus Q$_\beta$. Ese ácido nucleico se replica por mediación de un enzima específico que consta de cuatro subunidades. La cadena "más" del ARN vírico adopta habitualmente una configuración específica por apareamientos intracatenarios (*arriba, en el centro*). La cadena se despliega al avanzar la replicasa de su extremo 3' al 5' (*arriba, a la derecha*). El enzima va uniendo monómeros complementarios (nucleótidos precursores en forma activada), siguiendo las instrucciones del molde según las reglas de apareamiento, hasta fabricar una cadena "menos" complementaria (*abajo, en el centro*). La nueva cadena se pliega inmediatamente, impidiendo así la formación de una molécula bicatenaria, que detendría la replicación. La cadena "menos" se replica (*izquierda*) y forma una copia de la cadena "más" original. La autorreplicación por negativos es habitual en moléculas unicatenarias.

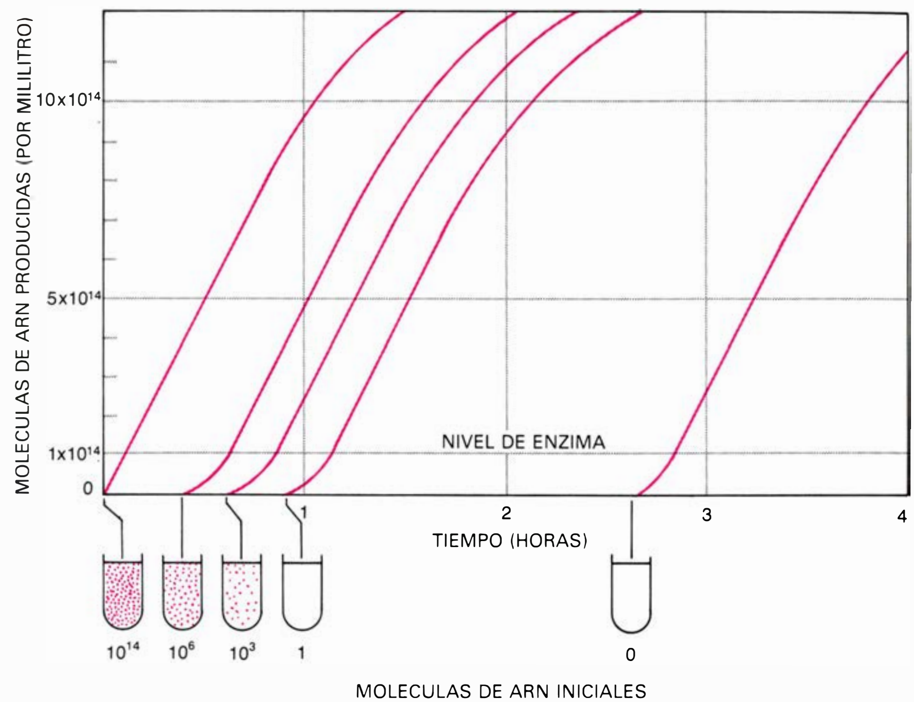


CICLOS INTERCONECTADOS de la síntesis de cadenas "más" y "menos"; son característicos de la replicación del ARN. El proceso tiene cuatro fases: iniciación de la replicación, elongación de la réplica, liberación de la réplica y reactivación. *E* es el enzima replicador; *I*, la molécula portadora de la información, o ARN molde; P_n es el producto, o ARN réplica, y *S*, el sustrato, o conjunto de monómeros. Las interacciones catalíticas regulan las concentraciones de cadenas "más" y "menos" de manera tal que la tasa de acumulación del conjunto de cadenas es la media geométrica de las tasas de cada clase.

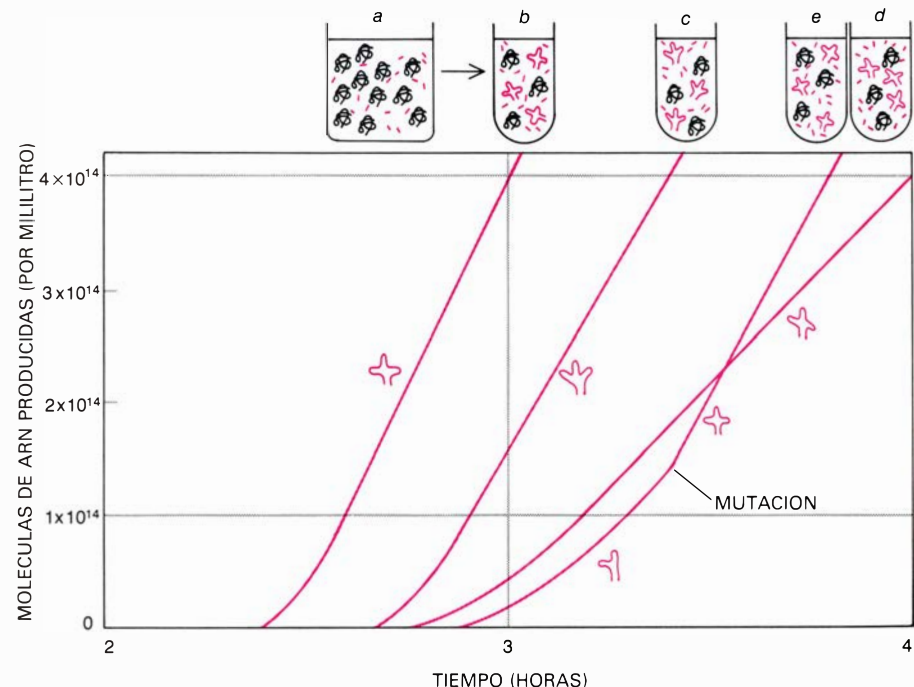
(*A*) y guanina (*G*) y las pirimidinas timina (*T*) y citosina (*C*); en el ARN, el uracilo (*U*) reemplaza a la timina. Las bases son complementarias y se aparean según reglas específicas: *A* con *T* (o *U*) y *G* con *C*. La complementariedad permite la replicación y la transcripción. En la replicación, una cadena de ADN o de ARN sirve de molde para disponer los nucleótidos complementarios según las reglas del apareamiento (con ayuda de varios enzimas llamados replicasas y polimerasas), formándose una cadena complementaria que contiene una copia de la información. En la transcripción, una secuencia de ADN origina, por un proceso similar, una cadena complementaria de ARN mensajero.

Conociendo las propiedades químicas del ADN y del ARN, ¿qué se puede deducir sobre la identidad de los primeros portadores prebióticos de información? Los nucleósidos de la desoxirribosa, componentes del ADN, son menos manejables químicamente que los de la ribosa, componentes del ARN. De hecho, la célula sintetiza los monómeros del ADN a través de intermediarios de ribosa; y la propia replicación del ADN se inicia con ARN. Los organismos actuales tratan su información genética con una complicada maquinaria de ARN y proteínas. Para que se haya originado tal maquinaria los propios portadores de información hubieron de ostentar características estructurales reconocibles. El ARN unicatenario se puede plegar en muchas estructuras tridimensionales, mientras que la doble hélice del ADN es uniforme. En las células actuales se encuentra ARN siempre que se requieren propiedades funcionales e informativas a la vez. No hay razón para pensar que fuera de otra manera en los primeros estadios de la vida, ni cabe imaginar que se haya transferido a ácidos nucleicos información almacenada de otra forma.

La búsqueda de la probable identidad química de los primeros genes conduce muy pronto al ARN. Podemos suponer que las primitivas rutas de síntesis y diferenciación proporcionaron bajísimas concentraciones de cortas secuencias de nucleótidos aceptables como "correctas" según la bioquímica actual: secuencias con las mismas bases, los mismos enlaces covalentes y la misma estereoquímica o disposición espacial de los grupos químicos. Estas secuencias coexistían, sin embargo, con miríadas de otras que serían tenidas hoy por "errores", con estereoquímica diferente, enlaces covalentes mal colo-



VELOCIDAD DE MULTIPLICACION de las cadenas de ARN. Se estudia incubando mezclas de cadenas de ARN molde de Q_β, monómeros y replicasa de Q_β. Este, Q_β, es un virus que infecta la bacteria *Escherichia coli*. Cuando hay más ARN molde que enzima, la acumulación de moléculas de ARN es lineal (hasta que las altas concentraciones del producto inhiben la acumulación). Este crecimiento es característico de las replicaciones sucesivas en las que se tiene que reactivar el enzima en cada replicación. Cuando hay más enzima que ARN molde, la acumulación es exponencial; tan pronto se completa un ciclo de replicación, la cadena nueva y la vieja se unen a moléculas libres de enzima y dan lugar a otras dos réplicas. Las curvas, que permanecen paralelas entre sí, se desplazan hacia la derecha conforme decrece la concentración inicial de ARN. Aun cuando no haya inicialmente molde, al cabo de un cierto tiempo acaba formándose, por interacciones entre el enzima y los monómeros, ARN *de novo* relacionado con fragmentos de Q_β.



SINTESIS *DE NOVO*. Se demuestra por experimentos de clonación. Una mezcla de monómeros y replicasas de Q_β libre de ARN (*a*) se incuba lo bastante para amplificar cualquier ARN que pudiera estar presente como impureza, pero no lo bastante para permitir la síntesis *de novo*. Se separa entonces la mezcla en cuatro tubos de ensayo (*b-e*), en cada uno de los cuales se acumulan cadenas de ARN nuevo a diversas velocidades (*curvas*). Un ARN molde óptimo surge por selección natural en cada tubo y acaba siendo el único producto; como la separación se hizo antes de la selección, los cuatro tubos contienen productos diferentes. A veces, una mutación aparece lo bastante tarde para ser vista (tubo *e*). A diferencia de la alta reproductibilidad de la replicación gobernada por un molde, se observan grandes fluctuaciones en el tiempo requerido para la aparición de productos *de novo*. Las fluctuaciones reflejan la participación de sólo unas pocas moléculas en el paso crucial, esto es, la síntesis de la primera cadena de ARN.

cados y bases atípicas. ¿Qué tenían de especial las secuencias parecidas al ARN actual?

Hay una contestación sencilla. Las cadenas de ARN con una estereoquímica homogénea y los enlaces covalentes apropiados se plegarían en estructuras secundarias estables, resultado de la formación de puentes de hidrógeno entre pares de nucleótidos complementarios. La ventaja sería importante, pues las haría más resistentes a la hidrólisis, esto es, a la rotura por agua, el sino final de los polímeros en solución acuosa.

Las cadenas primitivas de ARN que tuvieran el esqueleto y los nucleótidos apropiados gozarían de otra ventaja crucial: sólo ellas serían capaces de autorreplicarse establemente. Serían, a la vez, fuente de información (a través de las reglas de apareamiento) y productos a sintetizar siguiendo esa información. Encontramos aquí a nivel molecular las raíces del viejo problema del huevo y la gallina. ¿Qué fue primero, la función o la información? Como veremos, ninguna pudo preceder a la otra; tuvieron que evolucionar a la vez.

Las moléculas y los procesos prebióticos tenían seguramente muchos puntos comunes con la bioquímica actual. Sidney Fox y sus colegas, de la Universidad de Miami, han demostrado, por ejemplo, que los polímeros "protenoides", obtenidos, esencialmente, calentando una mezcla de aminoácidos (los constituyentes de las proteínas), pueden ejercer funciones enzimáticas. Además de tales catalizadores primitivos había sin duda moléculas capaces de activarse por la luz solar; había lípidos (grasas) o moléculas parecidas que podían formar estructuras membranosas y quizá también polisacáridos, o polímeros de azúcar, fuentes potenciales de energía. En breve, se formaron muchas moléculas funcionales por mecanismos químicos no vivos, o "no orgánicos".

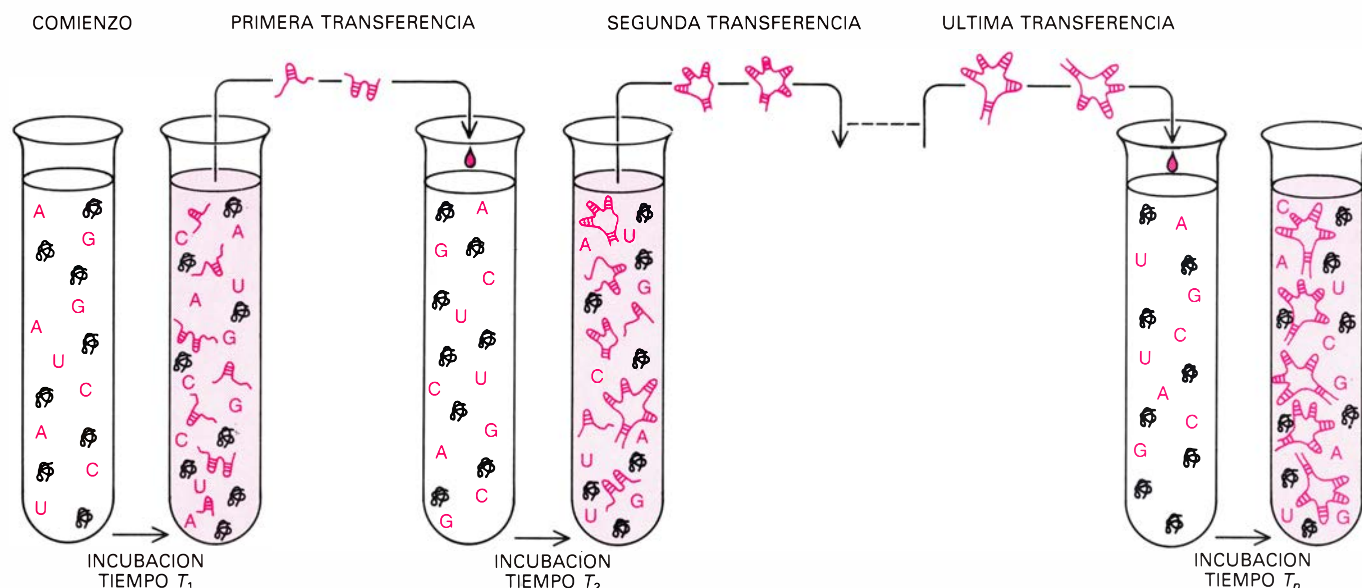
Tales moléculas funcionales pudieron ser importantes en la química de la sopa primitiva, pero no podían evolucionar. Su eficacia accidental dependía de condiciones estructurales no accidentales, tales como interacciones favorables con moléculas vecinas o determinados plegamientos en el espacio. Para llegar a ser más eficaces y para que pudieran seleccionarse las variantes más funcionales, tenían que escapar a tales condiciones estructurales. Sólo podrían lograrlo moléculas autorreplicables, capaces de conservar información. Examinaremos ahora el aumento de la complejidad y el contenido informativo de tales moléculas y la eliminación de las variantes menos funcionales.

Autorreplicación

El virus Q_β, que infecta a la bacteria *Escherichia coli,* sirve de modelo para estudiar la autorreplicación. Su genoma, o material hereditario, es una molécula de ARN unicatenario de unos 4500 nucleótidos. Sólo parte de esta molécula constituye el mensaje genético; el resto tiene varios papeles fundamentales (en vez de informativos), tales como el reconocimiento específico por ciertos enzimas. Hace varios años, Sol Spiegelman, a la sazón en la Universidad de Illinois, purificó la replicasa, o enzima replicador, de Q_β y demostró que reproducía el ARN del virus en un sistema experimental libre de células, originando copias infectivas. A partir de células de *E. coli* infectadas purificó también una molécula de ARN "satélite", no infecciosa, de 220 nucleótidos, replicada con extraordinaria eficacia por la replicasa de Q_β. El ARN satélite y otras "minivariantes" parecidas, en combinación con la replicasa de Q_β, sirvieron de sistemas apropiados para estudiar la replicación del ARN.

Un experimento típico comienza con una solución que incluye iones magnesio, una baja concentración de replicasa de Q_β muy purificada y una forma activada de los cuatro monómeros del ARN, los cuatro trifosfatos *ATP, GTP, UTP* y *CTP,* en los que la base y el azúcar están unidos a una cola de tres grugos fosfato. Para detectar la síntesis de ARN nuevo se marca con un isótopo radiactivo uno de los cuatro trifosfatos, de ordinario el *GTP.* Para iniciar la replicación se añade cierta cantidad de ARN molde y se incuba la mezcla.

Cuando Manfred Sumper hizo expe-



TRANSFERENCIA SERIADA, método diseñado por Sol Spiegelman, de la Facultad de Medicina y Cirugía de la Universidad de Columbia, para prolongar indefinidamente el crecimiento. Se empleó para demostrar la síntesis *de novo* de ARN y la evolución de los ARN óptimos. Preparó una serie de tubos de ensayo que contenían replicasa de Q_β, factores necesarios para el crecimiento y monómeros de adenina, guanina, citosina y uracilo, pero no moldes de ARN. Incubó la mezcla del primer tubo elevando su temperatura; tras un período de tiempo bastante prolongado se había sintetizado una mezcla heterogénea de cadenas cortas de ARN molde. Transfirió una pequeña fracción de esta mezcla al segundo tubo y la incubó. Repitió el proceso muchas veces. Se seleccionó, por fin, una única cadena óptima. Señal de la enorme multiplicación alcanzada por este método es que, si el crecimiento amplifica el producto unas diez mil veces en cada tubo, diez transferencias equivaldrían a un crecimiento que saturaría de ácido ribonucleico los océanos del mundo.

Modelo de las cuasiespecies

La química prebiótica del ARN proporcionó un ambiente apropiado para la evolución darwiniana: las poblaciones de moléculas autorreplicables (cadenas de ARN con distintas secuencias) competían por el "alimento" disponible (monómeros activados). La generación continua de secuencias mutantes, algunas de ellas ventajosas, obligó a reevaluar evolutivamente la especie óptima. Se ha desarrollado una teoría cuantitativa de esta competencia darwiniana.

Sea N_i el número de nucleótidos de la secuencia *i*. Identifiquemos la posición de cada nucleótido por un subíndice *p*, que puede valer desde 1 hasta N_i. Sea q_{ip} la probabilidad de que un nucleótido en la posición *p* de la secuencia *i* se copie correctamente durante la autorreplicación; la frecuencia de error para esa posición sería $1\text{-}q_{ip}$. El símbolo q_{ip} describe por tanto la exactitud o fidelidad de la copia al replicarse la posición *p* de la secuencia *i*. La probabilidad Q_i de que durante la replicación se forme una secuencia *i* totalmente correcta es el producto de las probabilidades para cada nucleótido:

$$Q_i = q_{i1} \times q_{i2} \times \ldots \times q_{iN_i} = \bar{q}_i^{N_i},$$

donde $\bar{q}_i$ es la medida geométrica de las fidelidades de copia de la secuencia *i*.

La secuencia *i* puede sobrevivir sucesivas replicaciones sólo si no se acumulan errores. Esto requiere que el crecimiento neto de la secuencia supere al de la media de sus competidores por un factor S_i. Además, sólo se puede seleccionar *i* si se satisface una condición de supervivencia: que la cantidad Q_iS_i, llamada umbral de error, sea mayor que 1.

El crecimiento neto está gobernado por la ecuación que describe las variaciones temporales de x_i, la proporción de todas las secuencias que son copias exactas de la secuencia *i*. Las principales causas de cambio de x_i son la replicación exacta de *i* y la replicación errónea de frecuencias muy parecidas, llamadas colectivamente *j*, que pueden originar *i* por mutación. Sumando ambas contribuciones obtenemos la tasa de cambio de x_i:

$$(W_{ii} - \bar{E})x_i + \text{suma de } W_{ij}\, x_j.$$

En esta expresión, W_{ii} es la tasa de replicación correcta de la secuencia *i*, y $\bar{E}$ es la tasa neta media de producción de todas las secuencias (diferencia entre las producidas por replicación y todas las pérdidas); ambas se expresan en tantos por copia. W_{ij} es la tasa de producción de secuencias *i* por copia errónea de secuencias *j*, dada en tantos por copia de la secuencia *j*. En la ecuación se suman las contribuciones procedentes de todas las secuencias llamadas *j*. Por tanto, el primer término es la tasa a la que compite la secuencia *i* con las demás secuencias y, el segundo, la velocidad de producción de las secuencias *i* por mutación de otras.

Esta expresión describe la evolución de cualquier conjunto arbitrario inicial de secuencias. El primer término será positivo o negativo según W_{ii} sea mayor o menor que la producción neta media $\bar{E}$. Si W_{ii} es mayor, x_i aumenta; si es menor, x_i disminuye hasta que la secuencia *i* desaparece o se produce sólo por mutación. La pérdida de secuencias con W_{ii} menor que $\bar{E}$ hace aumentar, sin embargo, el valor de $\bar{E}$, por lo que las secuencias supervivientes encuentran cada vez más difícil satisfacer el requerimiento de que W_{ii} sea mayor que $\bar{E}$. Estamos ante una especie de competición de salto de altura, en la que la barra va poniéndose cada vez más alta hasta que sólo queda un competidor, pero en la competencia molecular nunca queda un solo ganador, porque la mejor secuencia produce constantemente secuencias mutantes con las que tiene que seguir compitiendo (términos $W_{ij}\, x_j$). En el equilibrio dinámico que se alcanza, el mejor competidor, llamado secuencia maestra *m*, coexiste con todas las secuencias mutantes derivadas de él por error de copia. Llamamos cuasiespecie a esta distribución de secuencias.

Este análisis demuestra que el principio de la selección natural de Darwin no es un axioma, sino una consecuencia de las condiciones físicas de la autorreplicación. El resultado final de la selección, la cuasiespecie, es estable hasta que se produce por mutación una nueva secuencia más apta que la secuencia maestra entonces existente (o hasta que un cambio ambiental tenga el mismo efecto). Cuando esto ocurre, la nueva secuencia óptima prolifera hasta que predomina, acompañada por sus mutantes, y desaparece la cuasiespecie antigua.

Se han descrito cuantitativamente cuasiespecies de ARN. Por ejemplo, la secuencia maestra no puede tener más de

$$\frac{2{,}3 \log S_m}{1 - \bar{q}_m} \text{ nucleótidos.}$$

Las secuencias más largas no podrían sobrepasar el umbral de error, es decir, Q_mS_m no podría ser mayor que 1.

Este cuadro resume los principales resultados de nuestras investigaciones matemáticas y las posteriores de B. L. Jones, R. H. Enns y S. S. Ragnekar, de la Universidad Simon Fraser de la Columbia Británica, y de C. J. Thompson y J. L. McBride, de la Universidad de Melbourne.

rimentos de este tipo, en 1974, en nuestro laboratorio del Instituto Max Planck de Química Biofísica, en Göttingen, ocurrió algo totalmente inesperado. Al iniciar la incubación con más ARN molde que enzima, la concentración de ARN aumentó linealmente hasta alcanzar altos niveles. Esto nos indicó que todas las moléculas de enzima estaban ocupadas simultáneamente en la replicación de moldes; a pesar de que la concentración de moldes de ARN aumentaba continuamente, la concentración de complejos enzima-molde se mantenía constante; por lo tanto, se producía nuevo ARN a velocidad constante.

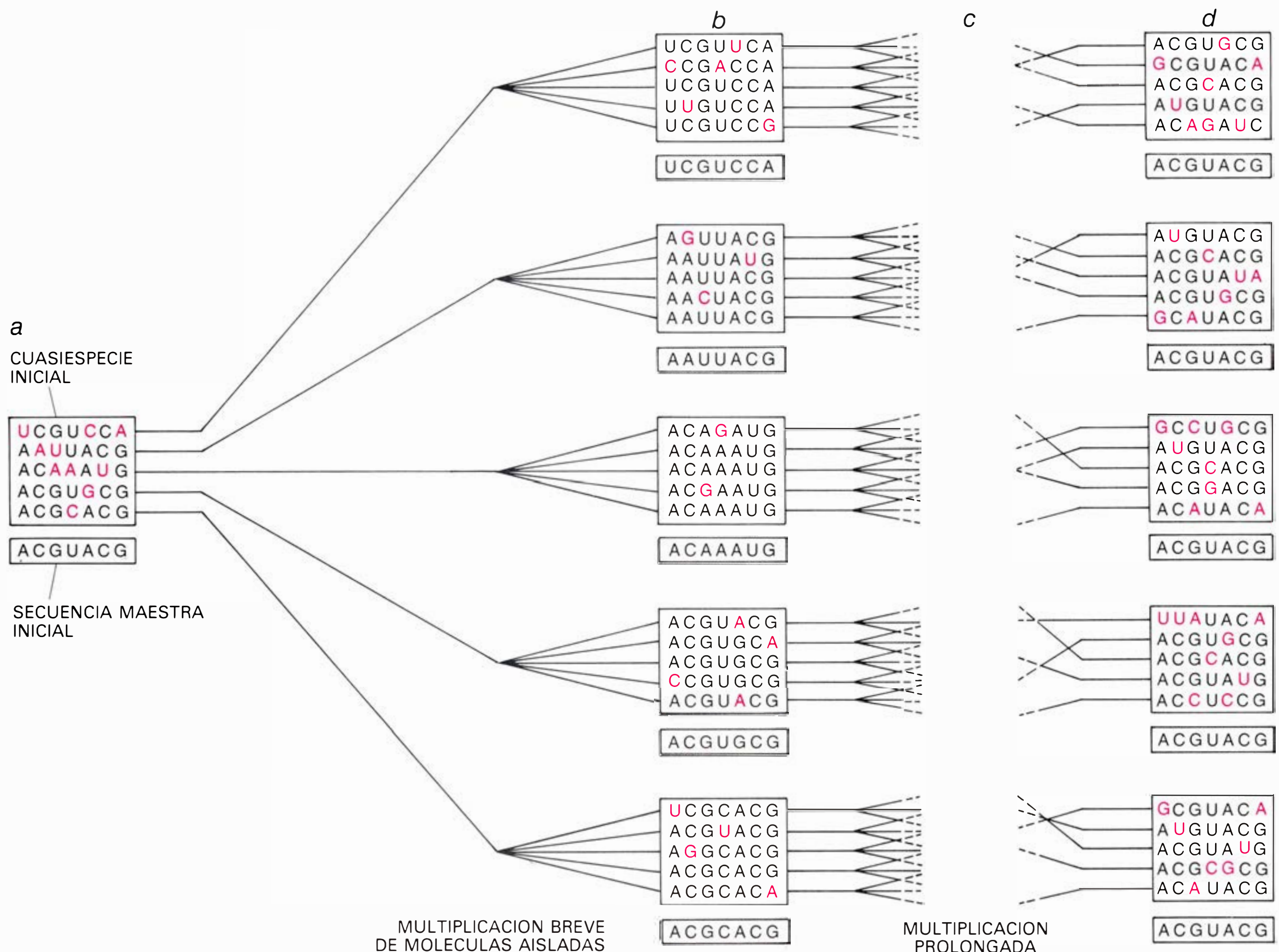
Era lógico repetir el experimento reduciendo la cantidad de ARN molde en la mezcla inicial. El resultado fue el desplazamiento en paralelo de la curva de acumulación lineal hacia un tiempo posterior [*véase la ilustración superior de la página 65*]. Las sucesivas reducciones retrasaron la acumulación proporcionalmente al logaritmo de la concentración inicial. En otras palabras, el paso de 10^6 a 10^4 moléculas de ARN por tubo de ensayo causó el mismo desplazamiento de la curva que el paso de

10^4 moléculas a 10^2. Esta relación logarítmica indicó claramente que, al sobrar enzima, toda molécula recién formada de ARN encontraba inmediatamente una molécula de enzima libre. La concentración de ARN aumentaba exponencialmente, no linealmente. El proceso funcionaba con pequeñísimas cantidades iniciales de ARN, aun con una sola molécula por tubo. (En este descubrimiento se basa un procedimiento para clonar una sola molécula.)

Imagínense nuestra sorpresa cuando Sumper descubrió que, aun sin añadir molécula alguna de ARN, seguía produciendose ARN, eso sí, tras tiempos de incubación mucho más largos y variables. Se eliminó por varios procedimientos la posibilidad de que las moléculas de enzima estuvieran contaminadas con ARN. Se sometieron los monómeros a condiciones en las que cualquier polímero se habría degradado totalmente. Se purificaron y analizaron los enzimas con todo el cuidado posible. Se añadieron deliberadamente impurezas para demostrar que causaban una forma de crecimiento completamente diferente. Finalmente, nos convencimos de haber obtenido moléculas de ARN sintetizadas *de novo* por la replicasa de Q_β. Lo más sorprendente era que el producto nuevo tenía siempre una composición parecida o idéntica a la minivariante de Spiegelman.

Los estudios comparativos de las velocidades de reacción evidenciaron pronto que los mecanismos de síntesis con o sin ARN iniciador eran muy diferentes. El complicado mecanismo de síntesis con molde de ARN se resolvió en pasos elementales, de modo que pudieran compararse cuantitativamente nuestras observaciones cinéticas con expresiones algebraicas. Una molécula de enzima se asocia con una molécula de ARN y fabrica una réplica, añadiéndose a cada paso un monómero de sustrato; no se observa cooperación entre los monómeros. Por el contrario, el paso limitante de la síntesis sin molde de ARN requiere la cooperación de tres o cuatro monómeros de sustrato por lo menos. En ese paso, además, participan al menos dos moléculas de enzima cargadas con monómeros. Una de las moléculas de enzima parece reemplazar al molde ausente exponiendo los monómeros que lleva al enzima polimerizador.

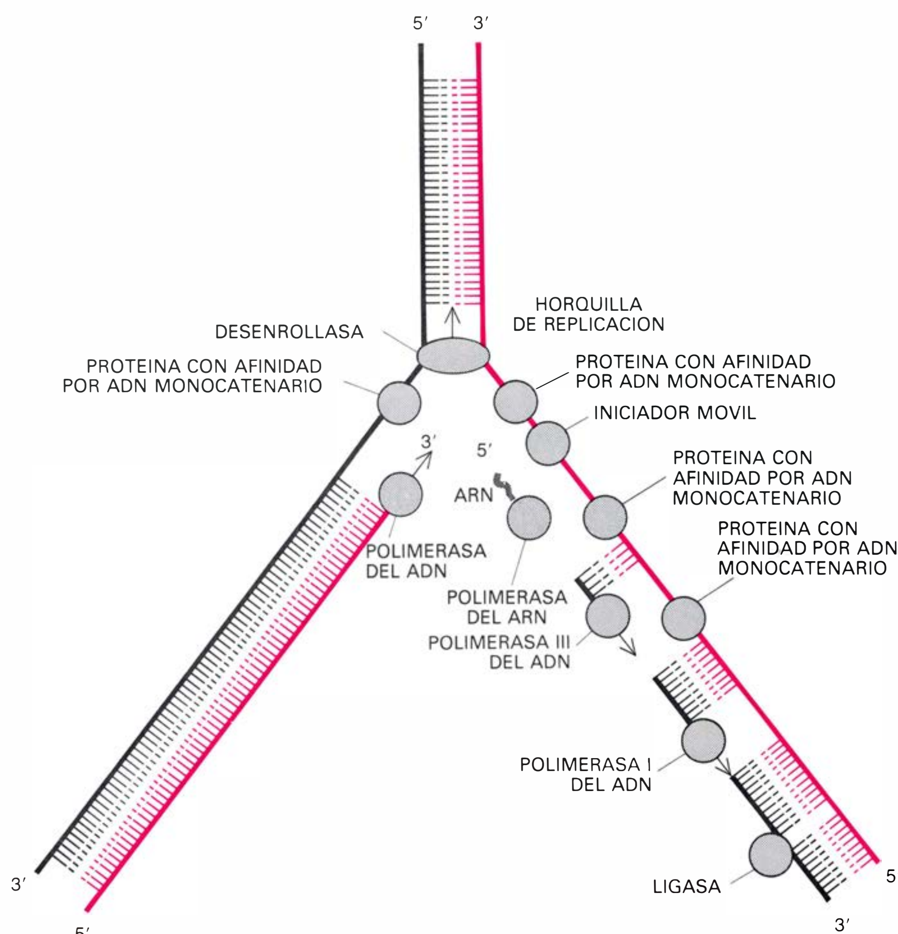
Spiegelman y Donald R. Mills, de la Facultad de Medicina y Cirugía de la Universidad de Columbia, han determinado la secuencia completa de la minivariante de 220 nucleótidos. Al analizar la secuencia observamos que podía



LA SECUENCIA MAESTRA UNICA de una cuasiespecie se mantiene a pesar de la continua aparición de secuencias mutantes, como demostró Charles Weissmann, de la Universidad de Zurich. Clonó el ARN de Q_β infectando bacterias con una solución tan diluida de la cuasiespecie original del virus (*a*) que cada infección partió de una sola partícula vírica. A continuación, analizó la secuencia de ARN de cada clon (*b*) por electroforesis bidimensional de ARN parcialmente fragmentado. Dejó que los clones se multiplicaran durante muchas generaciones (*c*), sometiéndolos así a una presión selectiva prolongada, lo que condujo al establecimiento de nuevas distribuciones de cuasiespecies (*d*). La ilustración esquematiza este experimento con cinco secuencias iniciales, cada una de siete nucleótidos. La secuencia maestra es la que presenta en cada posición el nucleótido más frecuente en esa posición. Se indican en color las diferencias respecto de la secuencia maestra. La secuencia maestra está siempre bien definida, aunque en realidad sea muy infrecuente. Después de la clonación, las nuevas secuencias maestras (*b*) son diferentes unas de otras y ninguna de ellas idéntica a la secuencia maestra original. Al cabo de muchas generaciones, sin embargo, todas las secuencias maestras (*d*) coinciden con la original. (La ilustración, como las demás del artículo, es de A. Beechel.)

representarse por repeticiones de cuatro tetrámeros y dos trímeros, aparte de 56 mutaciones y dos inserciones. Los tetrámeros eran *CCCC* y *UUCG* y sus complementarios, *GGGG* y *CGAA;* los trímeros *CCC* y su complementario *GGG* representan versiones truncadas de los tetrámeros. La secuencia *CCC* había sido identificada por Sumper y Bernd-Olaf Küppers como el lugar de reconocimiento que debe haber en cualquier ARN para interaccionar específicamente con la replicasa de Q_β; *UUCG* es la secuencia de bases que, en un contexto diferente (la traducción del ARN mensajero a proteína), se une con una de las proteínas que actúan de subunidades de la replicasa de Q_β.

¿Viola el descubrimiento de la síntesis *de novo* de ARN el dogma central de la biología molecular, según el cual la información sólo puede pasar de ácidos nucleicos a proteínas y no al revés? La selección de los tetrámeros y trímeros citados representa claramente una "consigna" por parte de las proteínas de la replicasa de Q_β. Sin embargo, los tetrámeros y trímeros citados podían haber compuesto muchísimas secuencias distintas, y no una sola. En los experimentos aparecen hasta 10^{12} moléculas molde, y sólo una de ellas tiene que ser amplificada. ¿No estaremos ante un caso de selección natural y no de información procedente de proteínas?



REPLICACION DEL ADN BICATENARIO, mucho más compleja que la del ARN. Incluye mecanismos para detectar y corregir errores en los que intervienen veinte o más enzimas. Un enzima desenrolla las dos cadenas parentales en la horquilla de replicación y ciertas proteínas con afinidad por el ADN unicatenario (o monocatenario) mantienen separadas las cadenas. Como las réplicas crecen siempre de 5' a 3', el proceso es discontinuo en una de las réplicas (*derecha*). Un iniciador móvil provee un lugar de reconocimiento para una polimerasa del ARN, que deposita un corto "cebo" de ARN (reemplazado más tarde por ADN). La polimerasa III prolonga el cebo con monómeros de ADN; la polimerasa I comprueba la secuencia, sustituyendo los nucleótidos incorrectos por correctos. Finalmente, la ligasa une los diversos fragmentos de la réplica. Si no hubiera corrección, la replicación del ADN no sería más precisa que la del ARN.

Papel de la selección

Esta pregunta halló respuesta recientemente en un experimento decisivo, realizado en nuestro laboratorio por Christof Biebricher y Rüdiger Luce, basado en la cinética especial de la síntesis *de novo*. Comenzaron incubando una mezcla sin ARN para amplificar cualquier impureza de ARN, pero sin dejar transcurrir el tiempo necesario que permitía la formación de ARN *de novo*. Separaron después la mezcla en varios tubos distintos y mantuvieron condiciones óptimas para la síntesis *de novo*. El resultado fue claro: aunque todos los tubos contenían una población uniforme de ARN, los ARN de distintos tubos eran diferentes; las diferentes secuencias, sin embargo, no carecían totalmente de relación.

El tiempo transcurrido hasta la aparición de ARN en tubos distintos fue muy variable. Estas fluctuaciones reflejaban la naturaleza probabilística de un proceso limitado por su primer paso, la síntesis de una sola molécula. Por el contrario, la amplificación de un molde inicial es determinista, con constantes temporales definidas, aun cuando la reacción comience con una sola molécula o muy pocas; las fluctuaciones en la velocidad de amplificación se compensan en replicaciones sucesivas.

Los primeros productos que aparecían en los tubos no habían sido optimizados todavía por ningún proceso evolutivo. Algunos tenían sólo unos sesenta nucleótidos, y durante las primeras etapas de la amplificación habían prevalecido probablemente otros más cortos. (Para analizar un ARN hacen falta por lo menos 10^{12} moléculas; este número equivale aproximadamente a 2^{40}, lo que implica que, antes de evaluar los productos, tienen que haber pasado unas cuarenta generaciones de amplificación, durante las cuales pueden haberse mejorado los moldes de ARN menos eficaces.)

Los experimentos de transferencias sucesivas, que permiten mantener el crecimiento durante muchas etapas de amplificación, nos dieron una idea de los productos óptimos. Generalmente tenían entre 150 y 250 nucleótidos de longitud. Cada conjunto de condiciones experimentales daba lugar a un producto final determinado, pero había tantos productos óptimos diferentes cuantas condiciones experimentales distintas. Uno de los productos óptimos resultó ser la minivariante de Spiegelman, que ya había aparecido sistemáticamente en las condiciones experimentales de Sumper. Otros productos óptimos estaban adaptados a condiciones que hubieran destruido la mayor parte de los ARN, tales como la presencia de altas concentraciones de ribonucleasa, enzima que fragmenta el ARN. La variante resistente a esta degradación parece plegarse de suerte tal que queden protegidos los lugares de posible corte. Otras variantes estaban tan bien adaptadas a ambientes extra-

ños que se podrían replicar en ellos mil veces más eficazmente que las variantes adaptadas a un ambiente normal.

Estas observaciones no dejan duda de la ocurrencia de síntesis *de novo* en los experimentos de Sumper. La uniformidad de los productos se interpreta como una consecuencia de la selección natural y no de información impuesta por el enzima. El dogma central queda a salvo, al menos en su esencia.

Más importe es lo que estos experimentos revelan sobre los procesos darwinianos. La selección natural y la evolución, consecuencias de la autorreproducción, actúan en el caso de moléculas como lo hacen en el caso de células o de especies. Lo realmente sorprendente, al par que descubrimiento de verdadera importancia, es la gran eficacia del proceso de adaptación en un sistema autorreproductor tan sencillo.

Molde sin enzima

Se puede objetar que un enzima como la replicasa de Q_β, una molécula tan compleja, no debería estar presente en un experimento que pretende representar la situación prebiótica, aunque no fuera sujeto de la evolución, sino un simple factor ambiental. La objeción, muy apropiada, nos conduce a otro problema importante.

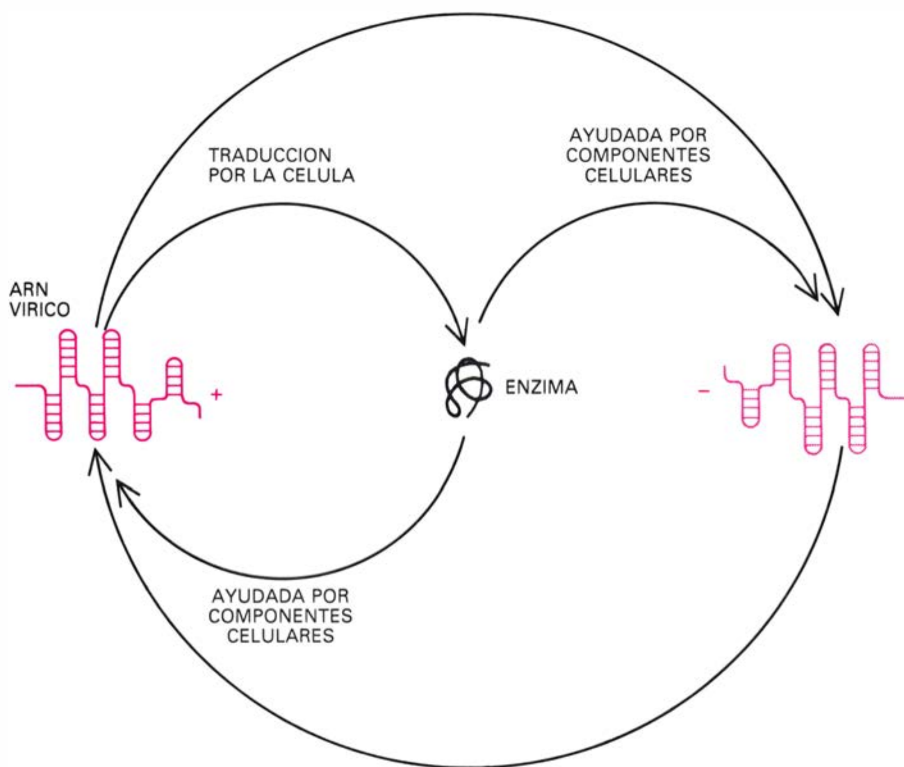
Si la replicación del ARN siempre hubiera requerido la participación de algo tan complicado como la replicasa de Q_β, la evolución prebiótica hubiera exigido otros procedimientos de optimización, además de la autorreproducción del ARN. Importa, por tanto, establecer qué clases de autorreproducción y de selección pueden ocurrir en ambientes sencillos que no incluyan replicasas bien adaptadas. Después podremos considerar el origen darwiniano de la síntesis de proteínas dirigida por información genética.

Este problema tiene que resolverse experimentalmente. El trabajo reciente de Leslie E. Orgel y sus colegas, del Instituto Salk de Estudios Biológicos, nos da pistas importantes. Se forman espontáneamente polímeros cortos del nucleótido adenina (oligo-*A*) al mezclar monómeros *A* con largos polímeros del nucleótido complementario *U* (poli-*U*), aunque no haya enzimas ni otros catalizadores. Las cadenas de oligo-*A* tienen una media de cinco nucleótidos y pueden llegar hasta diez. Si se añaden iones plomo como catalizador, el rendimiento mejora espectacularmente; además, la mayoría de los monómeros (75 por ciento) aparecen unidos, como en el ARN, por un grupo fosfato que forma un puente entre el carbono 3' de un azúcar y el carbono 5' del siguiente. Si una mezcla a partes iguales de monómeros *A* y *G* se incuba con poli-*C* e iones plomo, los productos tienen una proporción de 10*G*:1*A*, es decir, más del 90 por ciento de los apareamientos son correctos. En presencia de iones cinc y poli-*C*, los monómeros *G* forman cadenas de oligo-*G* de hasta unas 40 bases y la fidelidad es veinte veces mejor que con el catalizador de plomo. ¿Recuerda todavía la naturaleza cómo empezó la replicación? Todas las polimerasas actuales del ARN incluyen iones cinc.

Los resultados de Orgel indican que los polímeros ricos en *G* y *C* ofrecieron ventajas especiales a la evolución primitiva. Sólo ellos se copiaban con suficiente fidelidad en ausencia de replicasas apropiadas; sólo ellos producían apareamientos lo bastante fuertes como para que ARN mensajeros de un tamaño apreciable se tradujeran a proteínas activas en ausencia de ribosomas, sedes de la traducción en las células actuales. Los estudios cinéticos y termodinámicos de Dietmar Pörschke, en nuestro laboratorio, han proporcionado una base cuantitativa a estas conclusiones. El apareamiento *G-C* resulta ser unas diez veces más fuerte que el apareamiento *A-U*, de modo que las cadenas complementarias siguen unidas mucho más tiempo cuando abundan en *G* y *C*. Además, el enlace se ve reforzado cooperativamente por los apareamientos vecinos. De estos resultados hemos deducido reglas de apareamiento para un modelo evolutivo que permiten identificar estructuras bien conocidas del ARN (por ejemplo, la hoja de trébol de los ARN de transferencia) como el resultado evolutivo de procesos de ensayo y error.

La conclusión esencial de estos estudios sin enzimas es que el ARN se puede autorreplicar realmente sin ayuda de enzimas sofisticados. Podemos ocuparnos de las consecuencias evolutivas de la autorreplicación del ARN sabiendo que ocurrió realmente en tiempos prebióticos.

Admitamos un suministro inagotable de monómeros activados de ARN y atribuyamos vida eterna a las moléculas de ARN. ¿Qué clase de autorreplicación ocurriría? El ARN que se formara *de novo* serviría de molde y se reproduciría a una velocidad proporcio-



HALLAMOS HIPERCICLOS al nivel más elemental en los organismos cuando un virus de ARN infecta una célula. El virus asegura su replicación suministrando la información genética para un enzima que cataliza la multiplicación de su propia información. Se suministra ésta en forma de una cadena "más" de ARN, que la maquinaria celular del huésped traduce en un enzima. Con ayuda de factores del huésped, el enzima replica el ARN en una cadena "menos" que, a su vez, se replica en una nueva cadena "más". El doble circuito de retroalimentación, en el que la replicación del ARN depende de la secuencia del propio ARN y del enzima codificado por el propio ARN, equivale a una autocatálisis de segundo orden.

nal a su concentración. Resultaría un crecimiento exponencial. Aunque inicialmente sólo se hubiera formado una molécula, pronto habría muchas secuencias diferentes, porque en el curso de la replicación se cometerían inevitablemente errores, o mutaciones (sustituciones, inserciones y deleciones). En cada generación no sólo aumentaría el número de cadenas de ARN, sino también su variedad. ¿Qué ocurriría? Algunos mutantes se copiarían más deprisa que otros o serían menos susceptibles a errores al replicarse y sus concentraciones aumentarían más deprisa. Antes o después, estos mutantes predominarían.

Lo mismo ocurriría si los monómeros se suministraran sólo lentamente, de forma que los polímeros en crecimiento tuvieran que competir por ellos, o si se atribuyera duración finita a las cadenas de ARN. La autorreplicación es un proceso competitivo; el mejor competidor es la secuencia mutante con la combinación más favorable de estabilidad, fidelidad y rapidez en la replicación. Este es el hilo fundamental que hay que seguir para comprender los experimentos de autorreplicación que hemos descrito y nuestra teoría de la autorreplicación.

Cuasiespecies de ARN

El recuadro de la página 67 resume la teoría de la competencia en la autorreplicación molecular. El resultado de esta rivalidad es la "supervivencia" de la secuencia de ARN mejor adaptada a las condiciones existentes, a la que llamaremos secuencia maestra, acompañada de un séquito de secuencias similares derivadas por mutación de la secuencia maestra. Aunque desconocemos las constantes de acción de la química primitiva, podemos sacar conclusiones cuantitativas. Una de ellas es que hay una condición umbral para la autorreplicación estable de un mensaje genético. Hasta que las circunstancias permitieron que se cruzara ese umbral, no pudo sobrevivir ningún mensaje genético de ninguna clase.

Se pueden estimar las longitudes máximas de los genes disponibles en sistemas prebióticos insertando valores plausibles en la ecuación que determina la longitud de un gen en nuestro modelo de autorreplicación. Las longitudes máximas de los genes van de 50 a 100 nucleótidos, similares a las de los ARN de transferencia actuales. Constituye un resultado satisfactorio porque son suficientemente largos para el plegamiento interno y la estabilidad, pero en principio parecen mensajes genéticos demasiado cortos para determinar una proteína funcional.

Antes de mostrar la confirmación experimental de estos resultados teóricos, consideremos lo que se selecciona en la autorreplicación del ARN. En cierto sentido, se trata del gen más apto de los presentes (es decir, la secuencia maestra), porque esa secuencia es la más frecuente; pero la secuencia maestra constituye probablemente sólo una proporción pequeña del total de secuencias. En condiciones prebióticas, los mutantes serían extremadamente abundantes, porque la cinética química de la mayoría de las secuencias mutantes no debió de ser muy diferente de la cinética de la propia secuencia maestra. Por consiguiente, el producto resultante de la competencia en la autorreplicación sería la secuencia maestra y un gran número de mutantes derivados de ella, de los que no tendría forma de escapar.

Llamamos cuasiespecie a toda esta distribución de secuencias. Es la distribución de secuencias de la cuasiespecie lo que sobrevive a la competencia entre ARN autorreplicables, y no sólo la secuencia maestra o varias secuencias equivalentes, más aptas, de la distribución. La esencia de la selección estriba, pues, en la estabilidad de la cuasiespecie. Violar la condición sobre el umbral de error equivale a desestabilizar la cuasiespecie: la secuencia maestra no puede resistir la acumulación de errores; la distribución comienza a variar y acaba por perderse toda la información.

Las ecuaciones teóricas que describen la competencia durante la autorreplicación se han comprobado con experimentos de clonación y autorreplicación de Q_β realizados por Charles Weissmann y sus colegas, de la Universidad de Zurich. Midieron velocidades de replicación a corto y largo plazo y estudiaron la competencia entre clones mutantes y ARN de Q_β silvestre. Los valores de la fidelidad de la copia y de la ventaja competitiva obtenidos por análisis cuantitativo de los resultados estaban de acuerdo con la teoría. Los experimentos demostraron que, incluso con un sistema de replicación muy evolucionado, los organismos compensan la fidelidad imperfecta de la replicación limitando el tamaño de la información genética y manteniendo distribuciones de tipo cuasiespecie y no secuencias exclusivas.

Hemos mencionado la crisis energética que se tuvo que sobrepasar en las primeras etapas de la biogénesis. Analicemos ahora un obstáculo que desempeñó un papel aún mayor en la evolución de la vida: una crisis de información.

Error, genotipo y fenotipo

Los primeros sistemas moleculares darwinianos debían su capacidad de autorreplicación a fuerzas físicas inherentes que ocasionaban la formación de pares de bases complementarias. El umbral de error estableció un tamaño límite de unos 100 nucleótidos, alcanzable sólo por secuencias de ARN ricas en nucleótidos *G* y *C*. Ese límite pudo rebasarse cuando se desarrolló una capacidad para traducir genes a proteínas, obteniendo así una maquinaria enzimática que redujo la frecuencia de error lo bastante para permitir textos genéticos de hasta varios miles de nucleótidos. Este nuevo límite se refleja todavía en el tamaño de los actuales virus de ARN unicatenario, aunque los virus aparecieron mucho más tarde en la evolución.

Un tamaño aún mayor del texto genético sólo fue posible al aparecer mecanismos para detectar y corregir errores. Se podían reconocer los errores si la cadena hija seguía asociada a la parental, porque en ese caso podía detectarse químicamente el error en forma de desapareamiento.

Todo ello fue posible al aparecer en escena el ADN bicatenario. Las polimerasas del ADN corrigen la copia y suprimen errores tan eficazmente que permiten textos genéticos de hasta millones de nucleótidos. Lawrence A. Loeb, de la Facultad de Medicina de la Universidad de Washington, ha demostrado que, si una ADN polimerasa no puede corregir errores, tiene la misma fidelidad de replicación que una replicasa del ARN, o sea, entre 0,999 y 0,9999.

La invención del ADN posibilitó la aparición de células cuya división estuviera sincronizada con la replicación de su material genético. Pero se presentaba entonces una nueva crisis informativa: la replicación de alta fidelidad hizo más rara la aparición de nuevas variantes por medio de mutaciones puntuales. Esta dificultad fue compensada por el desarrollo de procesos de recombinación, entre ellos la reproducción sexual, que introdujo la genética mendeliana en la autorreproducción, base de los sistemas darwinianos.

De la primera crisis de información

sólo podía salirse a través de la organización de una maquinaria enzimática replicadora autoperfeccionable basada en una cuasiespecie estable. Este salto evolutivo requirió la traducción del texto del ARN en un nuevo texto: un texto funcional, el de las proteínas.

Para cifrar un aparato de traducción, por primitivo que fuera, se precisaba mucho más que el escaso centenar de nucleótidos que podían almacenarse reproduciblemente en una secuencia maestra. Comoquiera que apareciera la primera maquinaria enzimática, requería más información que la que podía proporcionar un sistema molecular darwiniano primitivo. Para estabilizar una cantidad mayor de información hizo falta la cooperación de genes diferentes, mediada y regulada por sus propios productos de traducción.

Si, en el nuevo sistema informativo, los productos de la traducción están sometidos también a evolución, surge otro problema. La selección tiene que actuar sobre el contenido informativo de la secuencia de nucleótidos: sobre el genotipo. Pero la evaluación de la selección ha de darse a nivel de la función del producto génico: el fenotipo. La dicotomía entre genotipo y fenotipo exige que el sistema permita llevar información a sus propios genes, procedimiento llamado autocatálisis de segundo orden. Es de segundo orden porque la reproducción de la molécula portadora de información necesita información suministrada por ella misma y por la maquinaria cifrada también en ella. Hemos llamado hiperciclos a los dobles circuitos de retroalimentación de este tipo. La expresión incluye un gran grupo de mecanismos autocatalíticos de orden superior. Su comportamiento en el tiempo difiere del de otros sistemas darwinianos.

Hiperciclos: las cuasiespecies cooperan

Un ejemplo actual de hiperciclo es la infección de una célula por un virus de ARN. Si el ARN vírico fuera sólo una molécula replicable más en el ambiente de la célula huésped, no lograría aventajar a los demás moldes del huésped. Pero lo que hace su información es especificar una maquinaria replicativa altamente selectiva para el propio ARN vírico. La mayoría de las piezas de esta maquinaria proceden del huésped, pero la interacción hipercíclica específica asegura el éxito del ataque vírico.

Un ejemplo sencillo explicará el principio operativo fundamental de un hiperciclo. Supongamos que una secuencia 1 de ARN cifra el enzima 1, que cataliza la autorreplicación de una secuencia 2 de ARN. La secuencia 2, a su vez, cifra un enzima 2, que cataliza la autorreplicación de la secuencia 1. ¿Qué sucede? La secuencia 1 necesita el enzima 2 para su autorreplicación y la secuencia 2 necesita el enzima 1. Por consiguiente, ninguna de ellas puede eliminar a la otra de la competencia por los monómeros disponibles; las dos secuencias tienen que cooperar. Según las velocidades de los numerosos pasos catalíticos, pueden prevalecer diversos niveles de concentración, pero, mientras haya interdependencia, sólo una fluctuación aleatoria excepcional y catastrófica o un cambio drástico en las condiciones químicas podrían extinguir

Modelo de los

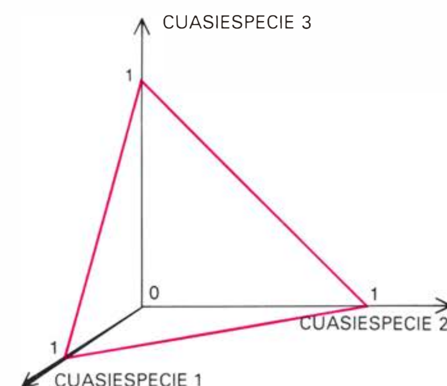
Un nuevo tipo de interacción dinámica molecular apareció en la evolución con la síntesis de proteínas dirigida por ARN. El análisis topológico, que alcanza conclusiones cualitativas, pero no cuantitativas, nos permite comprender sus características.

Consideremos un conjunto de varias secuencias maestras con sus mutantes acompañantes; cada secuencia maestra constituye con sus mutantes (en ausencia de las otras) una cuasiespecie estable. El contenido informativo total de todas las secuencias maestras rebasa el máximo permitido para una sola secuencia maestra por el umbral de error. Para que el conjunto permanezca estable y retenga el total de su información han de cumplirse tres condiciones: (1) cada cuasiespecie tiene que permanecer estable, es decir, cada secuencia maestra tiene que competir con sus mutantes de modo que no se acumulen errores; (2) las diferentes secuencias maestras, cada una con su propio valor selectivo, tienen que tolerarse unas a otras, forzadas por interacciones catalíticas mutuas; (3) el conjunto debe mantener estables las poblaciones de cada uno de sus miembros y competir con otros conjuntos similares.

El análisis topológico comienza definiendo un espacio en que cada eje de coordenadas representa la abundancia relativa de una cuasiespecie (el número de moléculas de ARN pertenecientes a esa cuasiespecie dividido por el total de moléculas de ARN). Para tres cuasiespecies se necesita un espacio tridimensional. Cada estado del sistema se caracteriza por las tres abundancias relativas y se representa por un punto en el espacio tridimensional. Como las tres abundancias relativas son positivas y suman la unidad, el punto tiene que caer en el triángulo equilátero cuyos vértices son el punto 1 de cada eje de coordenadas.

Llamaremos campo al triángulo equilátero. Si se consideran más de tres cuasiespecies, el campo sería otra figura geométrica en un espacio de más dimensiones. Los vértices del triángulo representan estados del sistema en que hay una sola cuasiespecie, los lados, presencia de dos cuasiespecies y, los puntos del interior, presencia de las tres cuasiespecies.

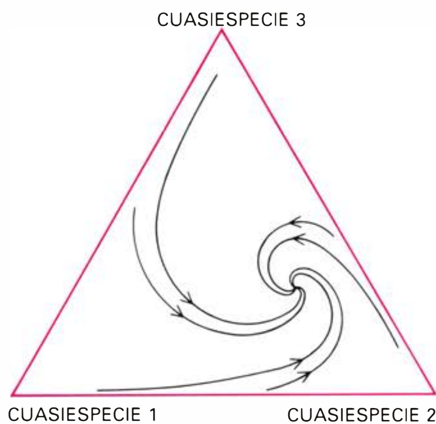
La variación temporal del estado de un sistema queda descrita por una curva dentro del campo, o "trayectoria". Hay métodos para averiguar la naturaleza cualitativa de las trayectorias sin resolver las ecuaciones dinámicas, que, si hay más de dos cuasiespecies, no se puede resolver analíticamente. Hay varias clases posibles de trayectorias. Cuando se da un equilibrio, las trayectorias convergen en un punto y, a partir de entonces, las abundancias relativas de las cuasiespecies permanecen constantes. Otra clase de trayectoria refleja oscilaciones de las poblaciones y, una tercera, un comportamiento irregular llamado caos.



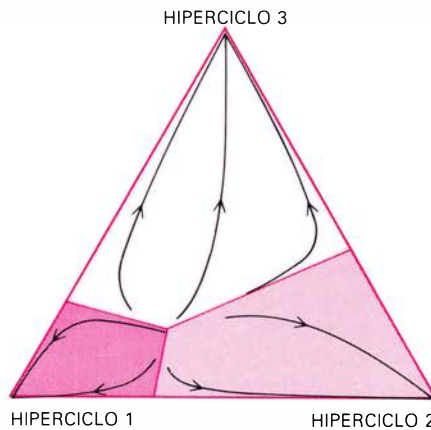
¿Qué formas de interacción hacen que las trayectorias permanezcan dentro del campo y coexistan, por tanto, todas las cuasiespecies? Una trayectoria que se dirija a un lado o a un vértice del triángulo significará la desaparición de una o más cuasiespecies. El análisis

hiperciclos

topológico general de los sistemas acoplados ha revelado que la coexistencia, y por tanto el cumplimiento de las tres condiciones anteriores, requiere un tipo particular de interacción, que hemos llamado hipercíclica. En un sistema hipercíclico, los ciclos de autorreplicación de las cuasiespecies están conectados por un circuito cerrado de interacciones catalíticas.



Los hiperciclos, como conjunto, tienen características dinámicas peculiares. La tasa de crecimiento de un hiperciclo no es proporcional a las poblaciones de cuasiespecies presentes, lo que conduciría a un crecimiento exponencial, sino a la población elevada a una potencia mayor que 1. Este crecimiento autocatalítico por encima del de primer orden se puede llamar hiperbólico. Los hiperciclos también difieren de los sistemas autorreplicativos darwinianos en que producen selección "de una vez para siempre". Una competencia entre hiperciclos se puede analizar topológicamente de la misma manera que una competencia entre cuasiespecies. Los ejes de coordenadas dan las abundancias relativas de los hiperciclos que compiten y las trayectorias caen en un campo de hiperciclos. Resulta que todas las trayectorias conducen al vértice del área del campo en que empezó la competencia, es decir, la competencia entre hiperciclos conduce siempre a la supervivencia de sólo uno de ellos.



La selección de una vez para siempre significa que un hiperciclo, una vez establecido, no puede ser desplazado por un competidor poco abundante, aunque sea más eficiente. Esto es consecuencia de que el valor selectivo de un hiperciclo depende de su población. No es ése el caso de los sistemas darwinianos, en los que una sola molécula mutante ventajosa llega a predominar sobre una población establecida. Los hiperciclos pueden evolucionar, sin embargo, optimizando sus relaciones internas como resultado de mutaciones en las moléculas portadoras de su información. Los hiperciclos crecen coherentemente porque las poblaciones que los constituyen se regulan mutuamente.

Los hiperciclos aparecieron en la evolución a la vez que la traducción del mensaje genético, que introdujo un nuevo requerimiento, el de revertir sobre los genes una evaluación de la calidad de los productos de su propia traducción. Probablemente, los hiperciclos siguen existiendo en las infecciones víricas.

el hiperciclo existente. En la autorreplicación intervienen moléculas replicables y enzimas, sirviendo el producto proteico de un ARN como replicasa, como activador de una replicasa o como otro elemento regulador que aumente la velocidad y la exactitud de la autorreplicación.

Se ha investigado detalladamente el comportamiento cinético de tales hiperciclos y se ha demostrado que constituyen las únicas redes funcionales que pueden rebasar el umbral de error de las cuasiespecies estables. El crecimiento hipercíclico es explosivo en comparación con un crecimiento autocatalítico de primer orden con constantes de acción parecidas. Las consecuencias del crecimiento hipercíclico sobre la selección son más peculiares.

La vida no se pudo originar con un sistema hipercíclico tan sencillo como el descrito. Los primeros circuitos catalíticos hubieron de ser débiles y complejos, con muchos participantes genéticos (miembros de la cuasiespecie de ARN) y funcionales (enzimas primitivos). Pero el principio hipercíclico era sencillo: la forzada cooperación entre genes, que de otra manera hubieran competido, permitió la mutua supervivencia y reguló su crecimiento. También posibilitó una evolución más refinada que la que podían realizar las cuasiespecies solas.

En las cuasiespecies, la competencia darwiniana evalúa la aptitud de cada ARN según su estabilidad y la velocidad y exactitud de su replicación. Al integrarse las cuasiespecies en un hiperciclo, aparecen nuevos criterios. En primer lugar, se hace crucial la evaluación de la función de cada cuasiespecie como sustrato: son más aptas las secuencias que se consiguen replicar más deprisa y con mayor fidelidad por el enzima responsable de su replicación. En segundo lugar, la introducción continua de nuevas secuencias mutantes significa que se ensayan incesantemente nuevas interacciones catalíticas. La estructura del hiperciclo evoluciona cuando las nuevas interacciones resultan ser ventajosas.

El hiperciclo comparte con la cuasiespecie una desventaja evolutiva. Tanto la competencia en cuasiespecies como la cooperación en hiperciclos evalúan sólo propiedades fenotípicas de los ARN: sus estabilidades y velocidades de reacción. Si se encontrara una forma de evaluar la calidad informativa de las secuencias de ARN, los enzimas que resultaran de la traducción de estas secuencias mejorarían por selección natural. Sólo se nos ocurre una forma: colocar las cuasiespecies organizadas en hiperciclos en compartimentos que puedan evolucionar por competencia darwiniana.

El paso de una cuasiespecie única a la organización hipercíclica de muchas cuasiespecies tuvo lugar probablemente de manera gradual, y no de repente. Los mecanismos primitivos de traducción de una cuasiespecie dieron lugar, antes o después, a proteínas más útiles para la autorreplicación que las proteínas aleatorias de la sopa primitiva. Al principio, la mejoría fue más o menos uniforme para toda la distribución de la cuasiespecie pero, al acentuarse las preferencias entre los productos de la traducción, la cooperación entre secuencias se hizo mucho más probable

que la mera relación entre cada secuencia y su producto. Las ventajas de una catálisis más específica fueron perfilando la complicada red inicial de interacciones. Finalmente, las diversas interacciones entre moldes y catalizadores se hicieron tan nítidas que cada enzima tenía su propio papel catalítico. Para entonces, la distribución original única de la cuasiespecie había divergido para formar un conjunto de distribuciones de cuasiespecies distintas y había entrado en funcionamiento el primer hiperciclo. Los hiperciclos surgieron tan natural y continuamente como las cuasiespecies; surgieron por ley natural.

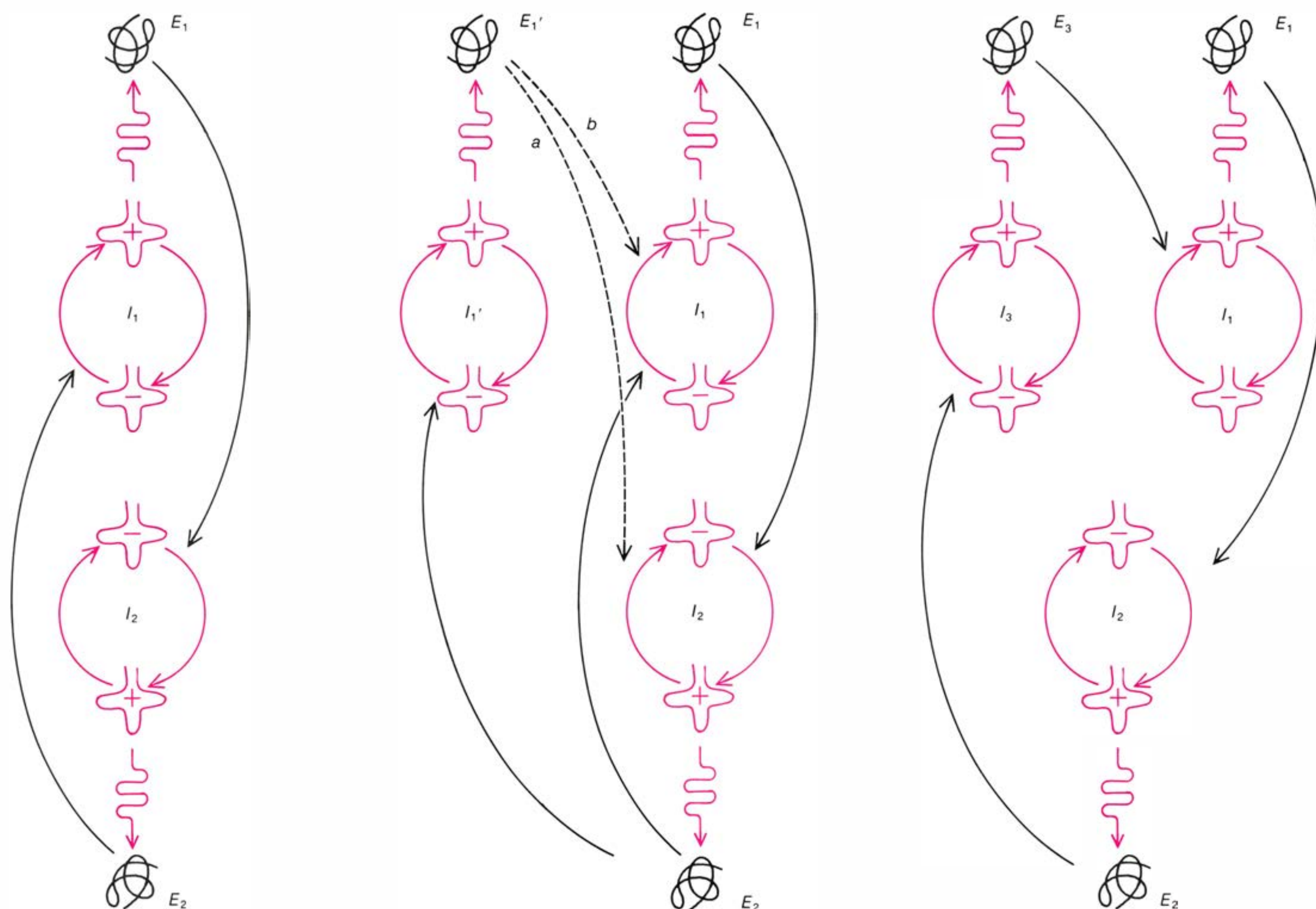
Compartimentación

La vida es ahora enteramente celular. ¿Por qué? Entre las obvias ventajas de la organización celular están la protección contra las fluctuaciones del ambiente externo y el mantenimiento de gradientes de concentración, pero tales ventajas no explican el origen de la organización espacial. Se necesitó la organización espacial porque constituía el único modo de resolver el problema informativo de la evolución que ni la competencia por la autorreplicación ni la cooperación hipercíclica eran capaces de resolver: la evolución del contenido informativo de los mensajes genéticos.

La organización celular se pospuso, sin duda, todo lo posible. Todo lo que impusiera límites espaciales en los sistemas homogéneos hubiera causado dificultades a la química prebiótica. La construcción de barreras, el transporte a su través y su modificación cuando hace falta son tareas realizadas hoy por los procesos celulares más refinados. Para alcanzar resultados análogos en una sopa prebiótica hubieran hecho falta innovaciones fundamentales.

La competencia darwiniana en una cuasiespecie se basó en la selección según la cinética química de las secuencias; el significado de las secuencias no importaba. No se pudo ignorar el significado de las secuencias cuando empezó a funcionar la organización hipercíclica de enzimas y ARN, porque del significado dependía su asociación. El carácter unidireccional de la interacción excluía todavía cualquier retroalimentación que permitiera evaluar la información genética, que permitiera seleccionar la mejor información. En un hiperciclo, como en una cuasiespecie, los ARN se evalúan sólo según su afinidad por las proteínas que los replican y no por la información genética que llevan. Un hiperciclo en solución no puede seleccionar sus productos de traducción, sean o no ventajosos.

Sólo vemos una posibilidad de evaluar la calidad de la información de los genes primitivos: destruir la homogeneidad de la sopa primitiva, compartimentar el proceso evolutivo. Al independizarse las mutaciones que ocurrieran en un compartimento de las que



ACOPLAMIENTO HIPERCICLICO. Permite la cooperación entre cadenas de ARN autorreplicables que, de otra manera, competirían entre sí y sólo permitirían la supervivencia de la más apta. En el hiperciclo de la izquierda, un ARN portador de información (I_1) cifra un enzima primitivo (E_1) que ayuda a replicarse a otro ARN (I_2) el cual, de manera similar, ayuda a replicar a I_1 a través de su producto (E_2). Este acoplamiento cíclico estabiliza las concentraciones de los ARN y los enzimas. En el hiperciclo central aparece una secuencia mutante (I_1') que compite ventajosamente con I_1; las consecuencias dependen de la función de su producto (E_1'). Si el nuevo enzima es más útil para I_2 que el original E_1 (*ruta a*) y si no tiene efecto sobre I_1, I_1' reemplazará a I_1 en el primer hiperciclo. Pero si E_1' es más útil que E_2 para I_1 (*ruta b*), el hiperciclo se ampliará a tres componentes (*derecha*).

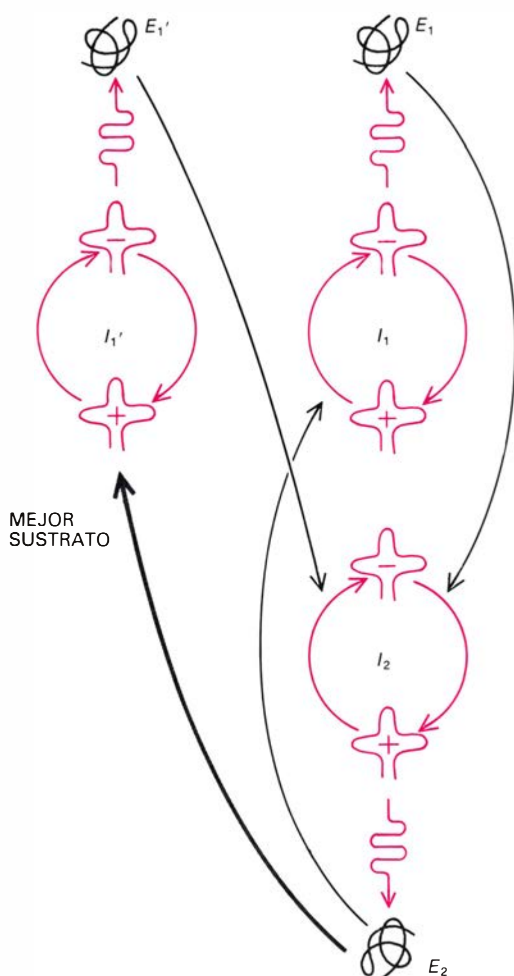
ocurrieran en otros, se había encontrado un medio de mejorar la información genética; ese medio fue, por supuesto, la evolución darwiniana. Los compartimentos más aptos pudieron seleccionarse según su funcionamiento global, incluyendo la posesión de mejor información genética. Mientras se pudiera transmitir información genética de una generación de compartimentos a la siguiente, se aseguraba la evolución del contenido informativo total.

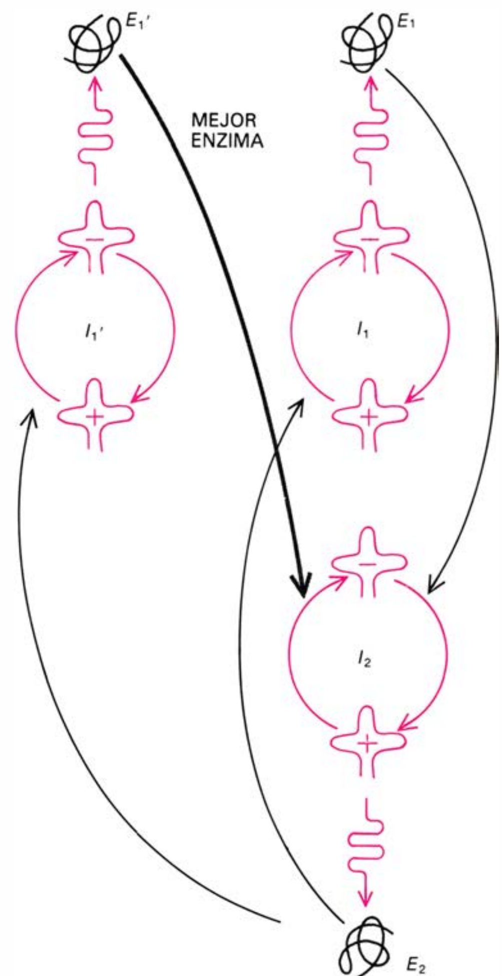
Con esto queda completa la estructura lógica de la evolución prebiótica. La estabilidad informativa requirió la autorreplicación de cortas secuencias de ARN. La competencia darwiniana entre secuencias mutantes condujo a la cuasiespecie como producto potencial de la evolución. Entre las secuencias mutantes se estableció después la organización hipercíclica, que permitió la coexistencia de muchas cuasiespecies en una misma sopa. La cantidad y variedad de información sobrepasaron con mucho lo que había sido posible con un solo gen primitivo (por su limitada fidelidad de copia), pero no se podía evaluar la información a partir de su función. Esta oportunidad de mejora evolutiva la suministró la compartimentación y la consiguiente competencia entre compartimentos.

Las exigencias lógicas del origen de la vida no se agotan con la compartimentación. En un compartimento se siguen dando los problemas de limitada fidelidad de copia y competencia entre los genes que se autorreplican. La organización hipercíclica es la única forma descubierta por ahora capaz de mantener suficiente información genética para cifrar las funciones enzimáticas mínimas necesarias. Los hiperciclos y los compartimentos son soluciones a dos problemas diferentes de la evolución prebiótica. Los hiperciclos permitieron la coexistencia estable de varios genes autorreplicables y resolvieron así la primera crisis informativa. Los compartimentos proporcionaron un modo de evaluar, y por tanto mejorar, el contenido informativo de los genes. En otras palabras, resolvieron la dicotomía genotipo-fenotipo.

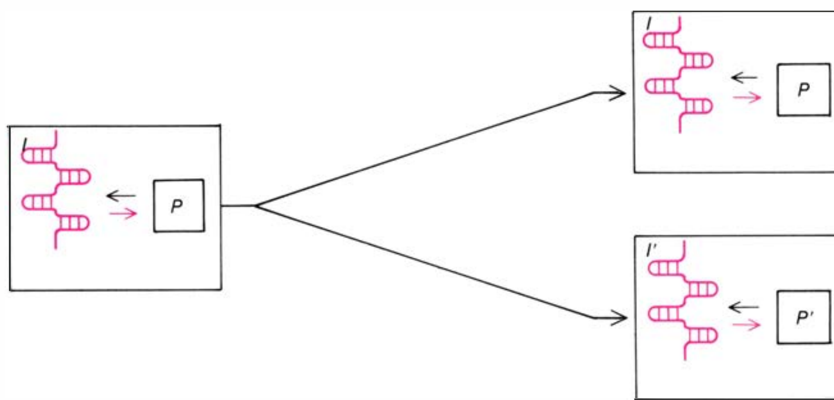
¿Vida en un tubo de ensayo?

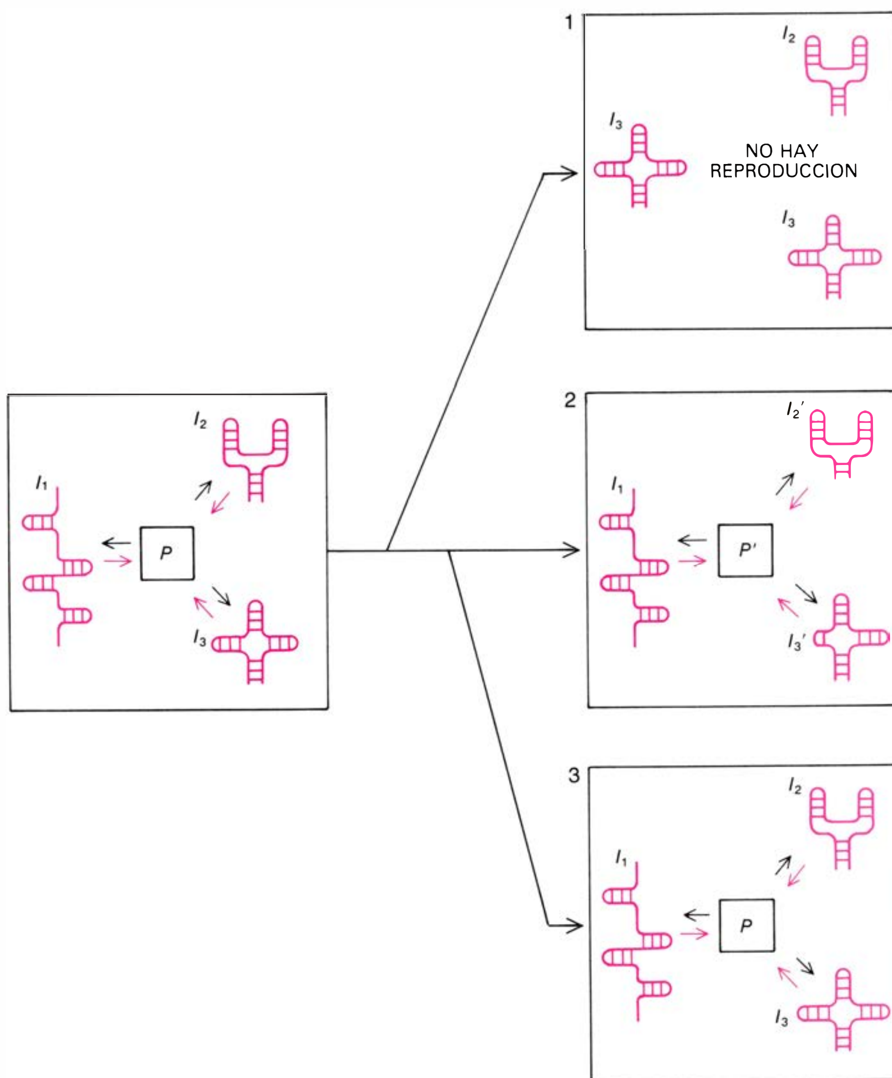
Si realmente pueden deducirse las leyes naturales que crearon la vida en la Tierra, ¿por qué no reunir los materiales apropiados y volver a crear la vida en un tubo de ensayo? El que intentara tal experimento subestimaría grandemente la complejidad de la evolución molecular prebiótica. Los investigadores sólo saben tocar melodías sencillas en uno o dos instrumentos de la gigantesca orquesta que interpreta la sinfonía de la evolución. El investigador reemplaza a un solo instrumento, de manera parecida a lo que hace un músico aficionado al acompañar a una grabación de la que se ha excluido a posta a uno de los músicos.





LAS MUTACIONES FENOTIPICAS Y GENOTIPICAS representaron papeles distintos en la evolución prebiótica. Decimos que una versión mutante (I_1') de la molécula informadora I_1 tiene un efecto fenotípico (*izquierda*) si es mejor sustrato del enzima E_2, esto es, si E_2 la replica más eficazmente que a I_1. El hiperciclo sobrevive si la nueva E_1' cataliza la replicación de I_2; si no, la mutación será "parásita": el mutante I_1' elimina a I_1 y el hiperciclo desaparece. Decimos, por otra parte, que un mutante de I_1 tiene un efecto genotípico (*derecha*) si su propia replicación por E_2 no cambia, pero el producto (I_1') desarrolla una actividad catalítica diferente de la de E_1 sobre la replicación de I_2. En una solución homogénea no habría selección en favor de la versión mejor de E_1. Tal evolución no fue posible hasta que los sistemas autorreplicables se aislaron espacialmente en compartimentos distintos.





LA COMPARTIMENTACION permite la selección de los genotipos. Supongamos que la molécula informativa *I* cifra la maquinaria (*P*) de la traducción y la replicación (*arriba*). Tras la mutación de *I* a *I'* y la traducción de *I'* a *P'*, el sistema se puede separar (*derecha*) en compartimentos hijos *I*/*P* e *I'*/*P'*. Esta separación permite la selección del gen que cifre la maquinaria más eficaz. Para que la selección ocurra, la superioridad de la maquinaria tiene que compensar los errores cometidos en cada replicación. Pero la compartimentación, por sí misma, no basta para seleccionar la información mejor, si ésta se reparte entre varias moléculas informativas (*abajo*). Los compartimentos hijos con componentes incompletos no son viables (*1*). La evolución es posible en conjuntos completos que incluyen mutaciones (*2*). Aunque el conjunto inicial sea el mejor (*3*), su estabilidad requiere una velocidad de proliferación mucho mayor que la requerida en una solución homogénea sin compartimentos. El efecto neto es un endurecimiento del umbral de error. Así, la compartimentación hace recaer sobre el genotipo la eficacia del fenotipo.

En este artículo, por ejemplo, hemos descrito algunos experimentos sobre evolución molecular que demuestran la formación de moléculas de ARN con propiedades fenotípicas muy refinadas cuando la maquinaria enzimática necesaria está presente como un factor ambiental. El paso siguiente sería descubrir el origen de tal maquinaria. Nos gustaría fabricar todos sus componentes y comprobarlos experimentalmente.

¿Qué aminoácidos formaron parte de las primeras proteínas sintetizadas bajo dirección de ARN? En la clave genética actual, cada uno de los 64, (4^3) tripletes posibles de los cuatro nucleótidos del ARN es un "codón" que determina la adición, por la maquinaria traductora, de uno de los veinte aminoácidos a la cadena de proteínas (y el principio y fin del proceso). Este sistema es, sin duda, demasiado complicado para que apareciera de golpe. ¿Hubo una clave más primitiva? ¿Cuál fue su estructura? Hay que considerar estas preguntas antes de diseñar experimentos sobre el origen de la traducción.

Las proteínas primitivas debieron estar compuestas de menos aminoácidos y una clave primitiva de uno o dos caracteres debió ser suficiente para determinarlas. Pero no hay un modo químico sencillo de cambiar una clave de uno o dos caracteres a una clave de tres, porque todos los mensajes preexistentes resultarían incomprensibles hasta que fueran reescritos por completo. Por tanto, la clave genética debió tener una estructura en tripletes desde el principio. ¿Cuál era esa estructura?

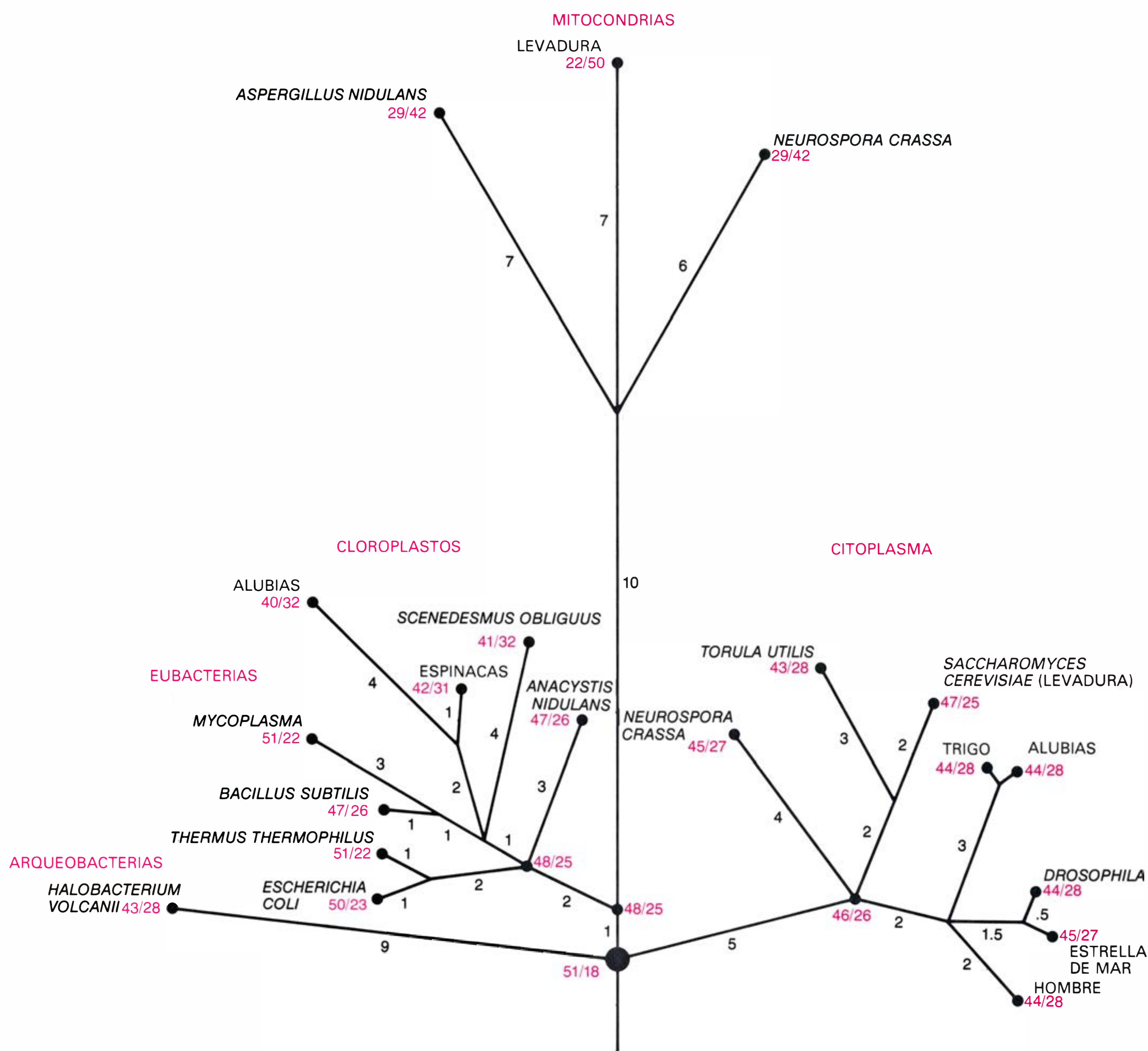
La traducción prebiótica impuso a la clave genética algunas tareas que el mecanismo de traducción actual resuelve con medidas refinadas, independientes de la propia clave. Al principio, la propia clave tenía que indicar la dirección de la lectura y la puntuación del mensaje definiendo una "fase de lectura". En 1976, F. H. C. Crick, Sidney Brenner, Aaron Klug y George Pieczenik, del Laboratorio de Biología Molecular del Consejo de Investigaciones Médicas de Cambridge, propusieron que inicialmente se fijó la dirección y la fase traduciendo sólo tripletes que tuvieran la secuencia *RRY; R* significa una purina (*G* o *A*) e *Y*, una pirimidina (*C* o *U*). Los tripletes *RNY* tendrían la misma función, con *N* igual a cualquier nucleótido. La dirección y la puntuación quedan establecidos igual de bien por *RNY* que por *RRY* y, con el primero, las dos cadenas complementarias poseen la misma estructura.

¿Se encuentran en la clave actual rastros de estructuras iniciales tales como *RRY* o *RNY*? Las secuencias archivadas en ordenador por Margaret Oakley Dayhoff y sus colaboradores, de la Fundación Nacional de Investigación Biomédica, han permitido realizar búsquedas a gran escala de relaciones genéticas entre polímeros biológicos y, en particular, construir árboles filogenéticos que dan la relación existente entre proteínas o ácidos nucleicos homólogos de especies diferentes. Los ARN de transferencia son particularmente apropiados para tales análisis en lo concerniente al problema de su origen. Su función consiste en acoplar cada aminoácido con su codón. Habida cuenta de ese papel crucial, cabe que su estructura refleje todavía la relación inicial entre codones y aminoácidos.

Evolución de la clave

Al buscar las secuencias primitivas analizando las presentes no bastó con tratar muchas secuencias con un programa de ordenador que obtuviera el árbol filogenético óptimo. Fue necesario formular y confirmar primero los criterios analíticos para determinar cuán "arborescentes" eran las secuencias, y esto se hizo mediante análisis topológicos llevados a cabo en cooperación con Andreas Dress, de la Universidad de Bielefeld.

Cuando se aplicaron estos criterios y programas como los de Dayhoff al análisis de todas las secuencias conocidas de ARN de transferencia (unas 200) surgieron dos conclusiones fascinantes. Primera: las secuencias de un ARN determinado (por ejemplo, el que media la iniciación de la traducción) de todas las especies estudiadas permiten construir un árbol filogenético que indica una divergencia evolutiva muy pequeña en comparación con la que se encuentra en otros biopolímeros. Al parecer, esta información primitiva se ha conservado bien en la evolución poste-



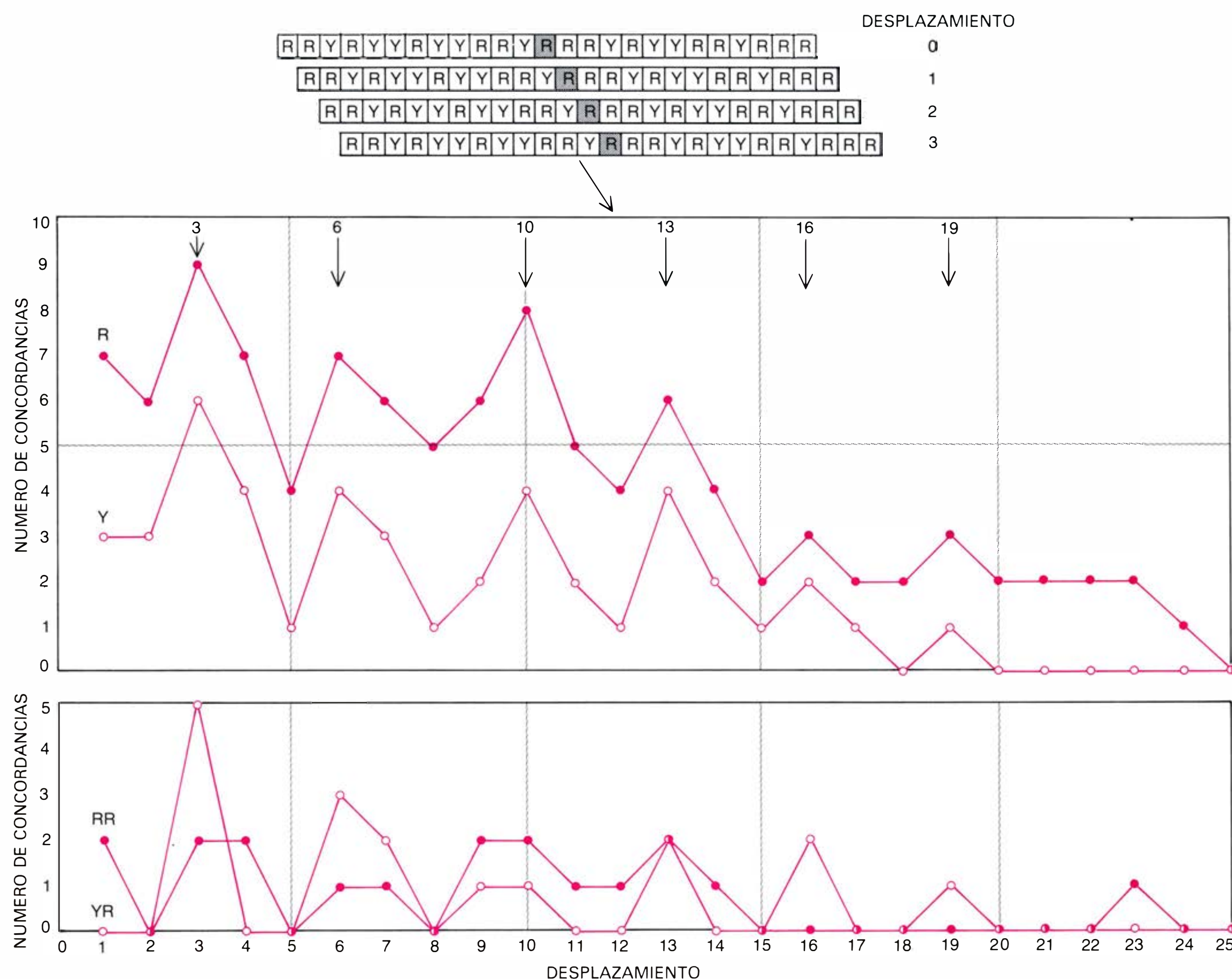
ARBOL FILOGENETICO del ARN de transferencia que reconoce el codón iniciador; revela que, en miles de millones de años, han ocurrido pocos cambios (*números negros*) en su secuencia de nucleótidos. La secuencia es casi la misma en todos los mamíferos estudiados; entre el hombre y la mosca *Drosophila* sólo hay unos pocos cambios. Los números en color dan la relación *GC/AU*. Esta relación es máxima (aproximadamente 2:1) cerca de los nodos primitivos del árbol y mínima en los extremos de las ramas largas (aproximadamente 1:2 en los ARN de transferencia de los orgánulos celulares llamados mitocondrias). Es decir, hoy no se precisa una relación alta; pudo ser necesaria inicialmente porque el apareamiento *G-C* es más estable que el *A-U*.

rior. Segunda: las secuencias de los diferentes ARN de transferencia de un mismo organismo indican un origen común, pero no parecen relacionarse unas con otras en un árbol genealógico, al menos no en los dos organismos (*E. coli* y levadura) de los que tenemos suficientes secuencias para que el análisis sea estadísticamente significativo.

Por el contrario, las secuencias parecen representar una distribución similar a la de los mutantes de una cuasiespecie. Al intentar remontarse a los tiempos del origen de la traducción, el análisis identificó lo que parecen ser los antepasados de los ARN de transferencia actuales y dio pie a dos importantes deducciones sobre ellos: tenían mucha más *G* y *C* que *A* y *U*, y sus secuencias maestras (determinadas asignando a cada posición la base más abudante en ella) mostraban una clara estructura de tripletes de la forma *RNY*.

También se puede buscar información genética antigua en otros lugares, dondequiera haya indicios de que la selección y la deriva genéticas no han amortiguado el "recuerdo" de las secuencias ancestrales por debajo del nivel de ruido. John Shepherd, de la Universidad de Basilea, aplicó recientemente un nuevo método de estudio de secuencias con ordenador, adecuado para largos mensajes genéticos. Su método mide la distancia entre repeticiones de caracteres o de grupos de caracteres en una secuencia. La información ancestral se puede distinguir de las modificaciones anteriores. Sus primeras conclusiones, sacadas de estudios de varios virus de ADN y genes de bacterias y organismos superiores, son que en estos genes modernos queda un recuerdo de las secuencias antiguas y que los tripletes *RNY* predominaron en tales secuencias.

La estabilidad de los apareamientos *G-C* sugiere fuertemente que la clave inicial *RNY* debió limitarse a los cuatro codones *GNC*. Las correspondencias actuales de estos codones son *GGC* = glicocola, *GCC* = alanina, *GAC* = aspartato y *GUC* = valina. Las simulaciones de la química primordial efectuadas por Stanley L. Miller, de la Universidad de San Diego, sugieren que estos aminoácidos estuvieron entre los más abundantes en la sopa primordial. Si es una coincidencia, resulta desde luego muy sugestiva.



ANALISIS ESTRUCTURAL DE GENES según un método de John Shepherd, de la Universidad de Basilea. La figura ilustra este análisis para una secuencia de codones *RNY* con la inserción de una base extra en la posición 13 y la sustitución de una *Y* por una *R* en la posición 25. *R* significa purina (adenina o guanina); *N*, cualquiera de las cuatro bases; *Y*, pirimidina (citosina o uracilo). Un programa de ordenador desplaza la secuencia una base a la derecha en cada ciclo (*arriba*) y cuenta el número de concordancias (en sentido vertical) de bases y pares *RR* e *YR* entre la secuencia desplazada y la anterior. Los máximos en el número de concordancias reflejan repeticiones dentro de la secuencia. Los máximos cada tres bases (con una perturbación causada por la inserción) están implícitos en una estructura de codones *RNY*. Se ha investigado de manera similar la estructura de secuencias naturales mucho más largas de ARN y ADN, prestando atención a posibles efectos extraños. La presencia de algunas correlaciones y la ausencia de otras se puede interpretar como indicios de antiguos codones *RNY* y confirman los argumentos teóricos en favor de tal codón y los resultados del estudio de los ARN de transferencia.

	PRIMERA BASE	SEGUNDA BASE G	SEGUNDA BASE C	SEGUNDA BASE A	SEGUNDA BASE U	TERCERA BASE
CLAVE *N* EN TRIPLETES *GNC, GNY*		GLY	ALA	ASP	VAL	

	PRIMERA BASE	G	C	A	U
CLAVE *RN* EN TRIPLETES *RNY*	G	GLY	ALA	ASP	VAL
	A	SER	THR	ASN?	ILE

	PRIMERA BASE	G	C	A	U	TERCERA BASE
CLAVE *RNN*	G	GLY	ALA	ASP	VAL	Y
		GLY	ALA	GLU	VAL	R
	A	SER	THR	ASN	ILE	Y
		ARG	THR	LYS	ILE/MET	R

	PRIMERA BASE	G	C	A	U	TERCERA BASE
CLAVE *NNN*	G	GLY	ALA	ASP	VAL	C
		GLY	ALA	ASP	VAL	U
		GLY	ALA	GLU	VAL	G
		GLY	ALA	GLU	VAL	A
	A	SER	THR	ASN	ILE	C
		SER	THR	ASN	ILE	U
		ARG	THR	LYS	MET ("INIC.")	G
		ARG	THR	LYS	ILE	A
	C	ARG	PRO	HIS	LEU	C
		ARG	PRO	HIS	LEU	U
		ARG	PRO	GLN	LEU	G
		ARG	PRO	GLN	LEU	A
	U	CYS	SER	TYR	PHE	C
		CYS	SER	TYR	PHE	U
		TRP	SER	"FIN"	LEU	G
		"FIN"	SER	"FIN"	LEU	A

EVOLUCION DE LA CLAVE GENETICA. Pudo haber comenzado con la correspondencia entre las cuatro bases *G*, *C*, *A* y *U* en la posición central del triplete y los cuatro aminoácidos entonces más abundantes. El triplete tenía la forma *GNC*, donde *N* es cualquier base, pues el apareamiento *G-C* es más fuerte que el *A-U*. Más tarde apareció *U* como alternativa de *C*, porque *G* a veces se aparea con *U*, originando tripletes *GNY* (*Y* es *C* o *U*). La presencia de *U* en tercera posición hizo que su complementaria, *A*, constituyera otra posibilidad para la primera posición, originando claves *RN* en una estructura *RNY* (*R* es *A* o *G*). La penetración de *R* en tercera posición, dando claves *RNN*, permitió que *Y* apareciera en primera posición en las cadenas complementarias y extendió el conjunto a los 64 tripletes actuales, *NNN*.

Hemos llegado a tener confianza en nuestra capacidad de reconstruir las secuencias ancestrales del ARN y las proteínas. A la vista de esta información estamos comenzando a reconstruir y resintetizar secuencias ancestrales, tanto de proteínas como de ARN y a comprobar su interacción en un reactor de flujo continuo, una especie de máquina evolutiva.

Si las primeras proteínas constaron realmente de los cuatro aminoácidos mencionados, tenían carga eléctrica negativa. En general, tales aminoácidos no se asociarían fácilmente con ARN cargado negativamente, a menos que otras fuerzas específicas estabilizaran una interacción particular. Claude Hélène, de la Universidad de Orleans, ha demostrado que hay una fuerte interacción específica entre los grupos carboxilo (COO^{-}) de aminoácidos como el aspartato y los nucleótidos *G* del ARN. En consecuencia, determinadas secuencias pueden facilitar contactos específicos, que se estabilizarían con ayuda de iones metálicos. Los primeros catalizadores específicos de la replicación y la traducción fueron probablemente estructuras de este tipo, que mediaban contactos específicos y ayudaban a funciones químicas débiles.

Todas estas funciones tuvieron que ser reclutadas de entre la información de una cuasiespecie inicial, cuyos mutantes se diferenciaron al asociarse en hiperciclos. Se han formulado y verificado experimentalmente los principios que guían la evolución de una organización de este tipo. Quedan todavía por describir las estructuras moleculares favorables.

Las envolturas de las novas

A diferencia de la supernova, la nova es una estrella enana blanca que expulsa una capa envolvente cuando una estrella compañera derrama sobre ella nuevo combustible nuclear. El espectro de la envoltura nos indica cómo se desarrolló

Robert E. Williams

La mayoría de las estrellas consumen su provisión de combustible nuclear a un ritmo notablemente constante durante cientos o miles de millones de años. Sin embargo, una vez cada diez años, por término medio, una estrella de nuestra región de la galaxia de la Vía Láctea aumenta bruscamente su brillo, multiplicándolo por un factor entre 10.000 y un millón y, durante un breve período, rivaliza con la más brillante estrella del cielo: se convierte en una nova. La nova brillante más reciente apareció en la constelación del Cisne a finales del verano de 1975.

No es fácil determinar la frecuencia total de novas en nuestra galaxia, que está formada por más de 100.000 millones de estrellas, distribuidas en un disco de unos 80.000 años-luz de diámetro. La dificultad obedece, principalmente, a la presencia del polvo interestelar. En ausencia de polvo, las novas más brillantes deberían ser teóricamente visibles a simple vista a una distancia de 25.000 años-luz. Sin embargo, en una región altamente cargada de polvo, una nova típica podría fulgurar a distancia tan cercana como 1000 años-luz sin ser jamás observada. La mejor estima de la frecuencia con que se producen las novas se ha obtenido vigilando su aparición en galaxias espirales próximas. Tales exploraciones han demostrado también que se pueden distinguir dos tipos cualitativamente diferentes de explosiones estelares: novas y supernovas. En una galaxia típica se dan unas 25 novas por año y sólo dos o tres supernovas por siglo. En nuestra galaxia no se ha visto ninguna supernova desde 1604. Las supernovas son decenas de miles de veces más brillantes y más energéticas que las novas.

Ambos fenómenos, completamente diferentes, no están relacionados entre sí. Una supernova representa la breve culminación final en la evolución de una estrella de gran masa. A temperaturas y presiones suficientemente altas, los núcleos de los elementos pesados de la parte central de la estrella participan en una serie de reacciones que extraen energía de la estrella, haciendo que ésta se contraiga rápidamente para después expansionarse de modo explosivo. En el proceso, la estrella se desmembra, con excepción de un núcleo central de gran densidad, que es lo que queda como residuo en la forma de una estrella de neutrones. Las novas, por otro lado, no son parte de la evolución de las estrellas normales. Son episodios termonucleares que tienen lugar, probablemente de manera periódica, en la superficie de una estrella enana blanca que está estrechamente ligada, mediante fuerzas gravitatorias, a una estrella en expansión, mucho mayor y más fría.

A pesar de sus diferencias básicas, las novas y supernovas tienen un rasgo común: ambas expulsan materia gaseosa al espacio. Los restos de explosiones de supernova que configuran nebulosas filiformes pueden persistir siglos enteros. La Nebulosa del Cangrejo, en Tauro, es el resto de una supernova que explotó en el año 1054 de la era cristiana. Las envolturas expulsadas por las novas son más pequeñas y de menor masa que los residuos de supernovas y, generalmente, se pueden ver sólo alrededor de las novas más próximas y brillantes. El reciente examen espectroscópico de las capas envolventes de las novas ha suministrado datos sobre los tipos de reacciones termonucleares que constituyen el motivo de la explosiva liberación de energía en las novas. Estos estudios han mostrado que las envolturas de las novas son más ricas que las estrellas normales en elementos pesados tales como el carbono, nitrógeno y oxígeno.

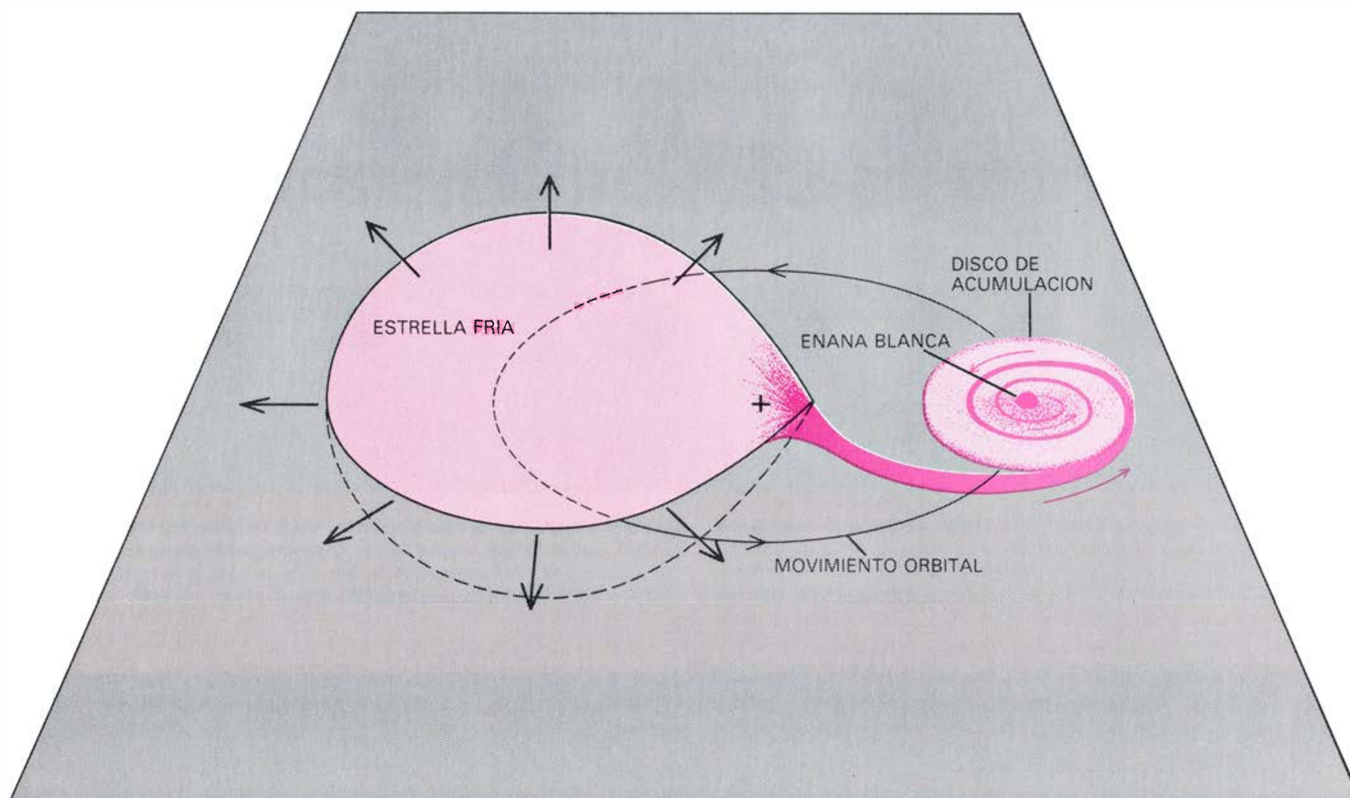
El tiempo requerido por una nova para alcanzar su máxima luminosidad puede variar considerablemente: desde un par de días hasta varios meses. La gráfica del brillo visual de una nova a lo largo de un cierto período de tiempo se llama la curva de luz de la nova. Las curvas de luz sirven para distinguir dos clases generales de novas: rápidas y lentas. Las novas rápidas, generalmente, multiplican su brillo por más de 10.000 en sólo unos pocos días. Su máximo brillo no dura ni siquiera una semana, para decrecer luego continuamente. En un comienzo, el decrecimiento de una nova rápida es bastante veloz: del orden de hasta una magnitud estelar (un divisor de 2,512) cada dos días. Las novas lentas alcanzan su máximo brillo más gradual e irregularmente que las novas rápidas, y declinan mucho más lentamente. Además, su aumento de brillo es menor. Sin embargo, la energía total liberada en la explosión viene a coincidir en ambas clases de novas. Por lo que hasta ahora se sabe, las novas recobran con el tiempo el brillo que tenían antes de la explosión.

Se designan, inicialmente, según la constelación en la que se han observado y el año de explosión. Así, la nova de 1975, del Cisne, que a lo largo de varias noches brilló tanto como la estrella vecina Deneb, de primera magnitud, se designó Nova del Cisne 1975 (o Nova Cygni 1975). Posteriormente, se da a la nova un nombre oficial: al genitivo latino de la constelación se añade un prefijo formado por letras (o V, para indicar variable) y un número que designa el orden del descubrimiento o de las estrellas variables en aquella constelación. De aquí que la Nova del Cisne 1975 se conozca ahora oficialmente como V1500 Cygni, y otra nova brillante bien conocida, la Nova de Hércules 1934, responda a la denominación de DQ Herculis.

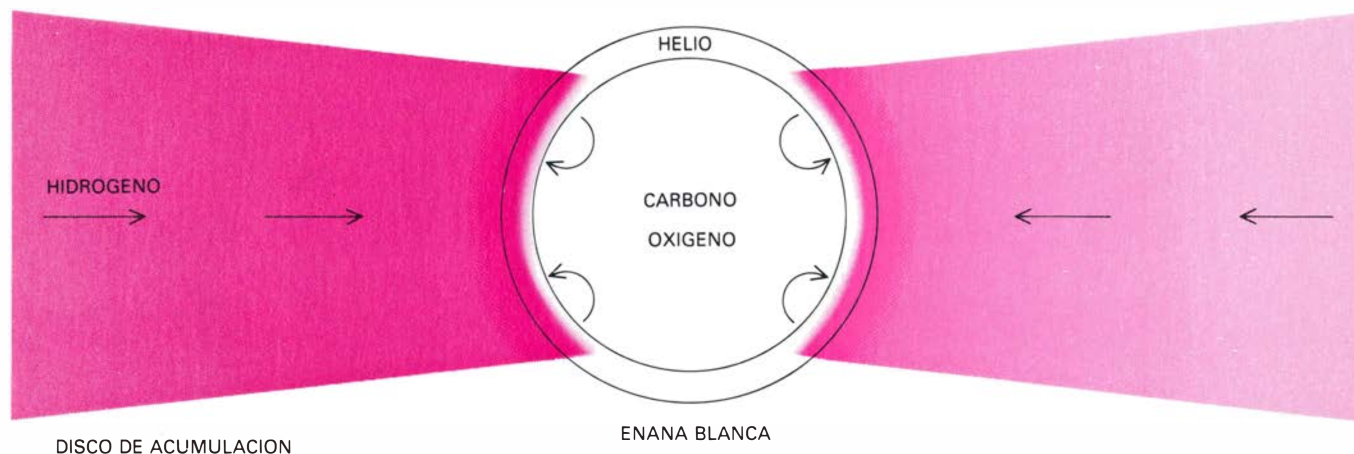
Nuestro conocimiento de las explosiones de nova arranca del descubrimiento, en 1954, por Merle F. Walker, de los Observatorios del Monte Wilson

y Monte Palomar, de que DQ Herculis era una binaria eclipsante, o sistema de doble estrella. Walker estaba siguiendo la emisión de luz de DQ Herculis, en su intento por determinar la naturaleza del parpadeo que mostraban muchas novas antiguas, cuando observó que el brillo de la estrella decrecía ostensiblemente durante casi una hora y volvía después a su intensidad original. Observaciones ulteriores establecieron que el proceso se repetía cada cuatro horas 39 minutos: sin duda, una compañera invisible que giraba a su alrededor estaba eclipsando la estrella. El descubrimiento de que DQ Herculis era en realidad una estrella doble, con el período de revolución más corto de todos los sistemas binarios conocidos hasta entonces, permitió determinar algunas propiedades fundamentales de las dos estrellas.

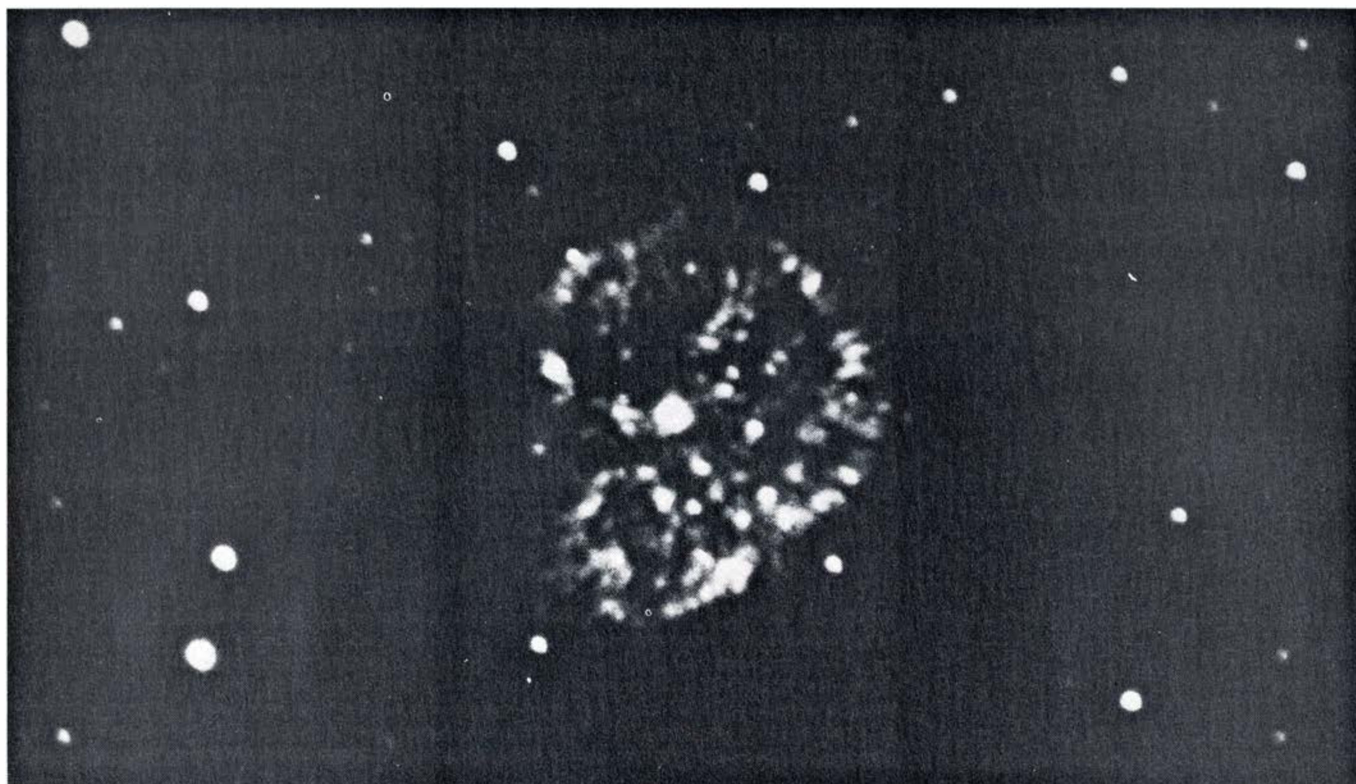
Una ley bien conocida en astronomía, que puede deducirse de las leyes newtonianas sobre el movimiento y la gravitación, establece que la separación entre dos objetos ligados gravita-



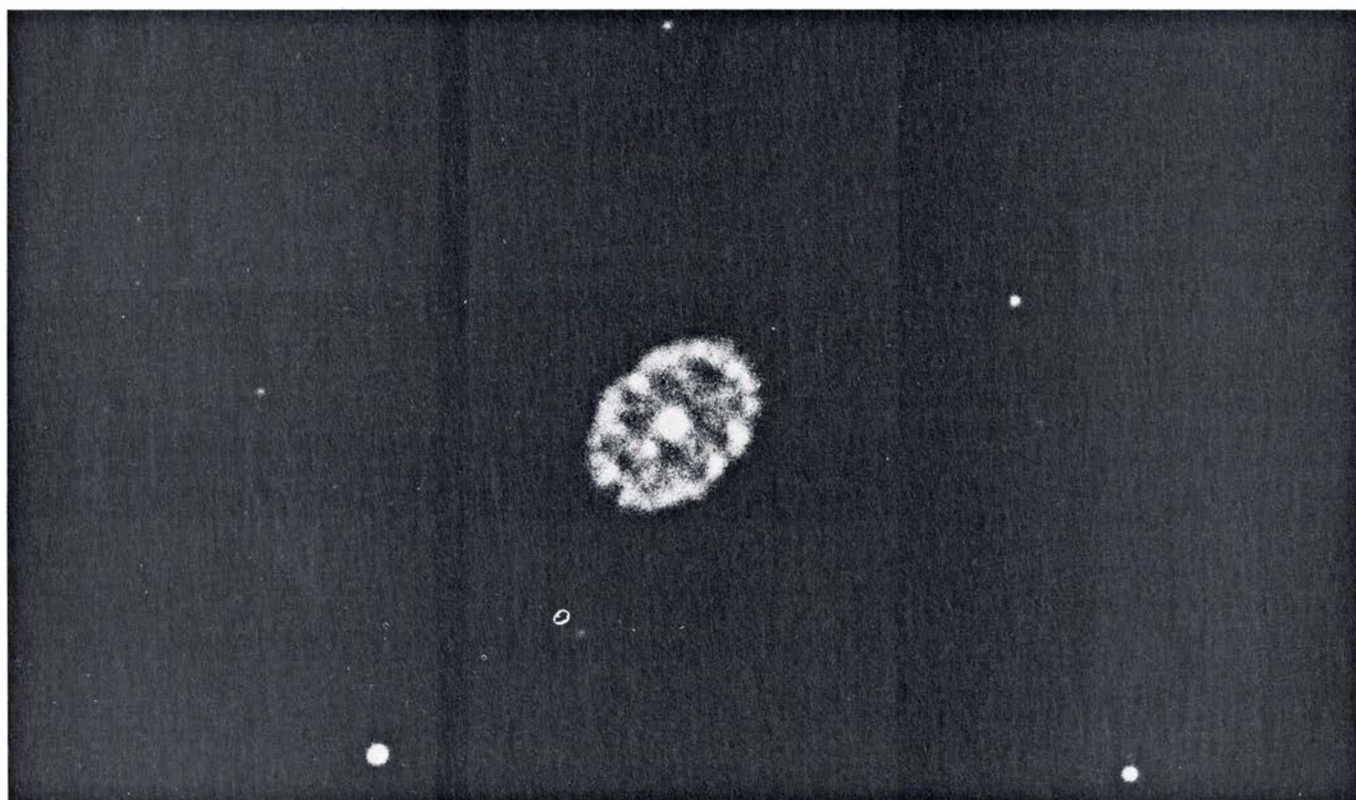
SISTEMA BINARIO DE NOVA representado esquemáticamente desde encima del plano de las órbitas. La cruz es el centro de rotación. El miembro enana blanca del sistema es una estrella que ha convertido su provisión inicial de hidrógeno y helio en elementos más pesados. Careciendo de fuente de energía, se ha contraído hasta un diámetro como el de la Tierra. Su compañera es una estrella de tamaño normal que ha agotado la mayor parte del hidrógeno en su zona central y ha comenzado a expandirse a medida que las reacciones termonucleares se van propagando desde el interior hasta la superficie. Al dilatarse la estrella en su evolución hacia la clase de gigante roja, su forma se altera en virtud de la atracción gravitatoria de la enana blanca. El gas hidrógeno que se desprende es arrastrado hacia un disco de acumulación que gira alrededor de la estrella más pequeña. En la fase final, la mayor parte del gas cae sobre la enana blanca tras describir espirales a gran velocidad, elevando la temperatura de la superficie de la enana y preparando así la explosión.



EL CHOQUE DEL GAS a gran velocidad procedente del disco de acumulación, representado aquí "de canto", da lugar a una capa periférica caliente en la superficie de la enana blanca. La materia altamente comprimida en la enana blanca se halla en un estado degenerado: se parece más a un sólido que a un gas. Por tanto, no se dilata al calentarse. El hidrógeno del disco de acumulación se mezcla con elementos más pesados, como el carbono y el oxígeno, que pueden ser transportados a la superficie de la enana blanca, desde el interior, por corrientes generadas por el gas incidente. A la temperatura crítica, 20×10^6 grados, la fusión del hidrógeno y el carbono inicia la cadena de reacciones nucleares controladas que culminan en la explosión de nova.



ENVOLTURA DE GAS EN EXPANSION, visible alrededor de la nova GK Persei, de 1901. La nova alcanzó su máximo rápidamente, brillando por breve tiempo más que Procyon, la octava estrella del cielo en cuanto a luminosidad. La envoltura no sólo es la mayor de cuantas se han observado, sino que también resalta por ser heterogénea y muy caliente. Su temperatura, de 30.000 grados Kelvin, supera la que una nova puede mantener. El aspecto apedazado y su alta temperatura le vienen del choque de la envoltura, que se dilata a 3000 kilómetros por segundo, con el gas en el medio interestelar.

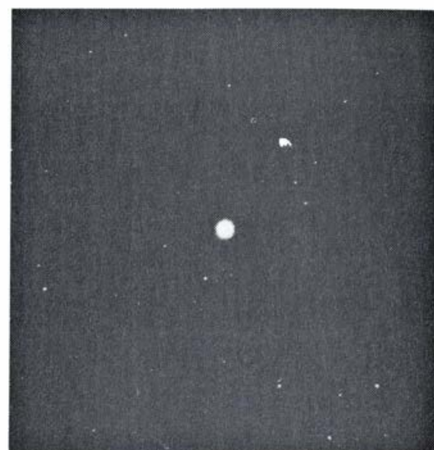


ENVOLTURA ALREDEDOR DE DQ HERCULIS, expulsada por una nova lenta que tardó semanas en alcanzar su máxima intensidad en 1934, cuando llegó a rivalizar en brillo con Deneb, estrella en la cercana constelación del Cisne, cuyo brillo la sitúa en el 19º lugar en el cielo. La fotografía, impresionada por la luz emitida por el hidrógeno, se tomó con el telescopio de 2,3 metros del Observatorio Steward. A diferencia de la envoltura alrededor de GK Persei, la que ciñe DQ Herculis es simétrica. Aunque el gas de la envoltura se encuentra ionizado, como si estuviera caliente, en realidad está muy frío, a 500 grados Kelvin. El análisis espectral de la envoltura realizado por el autor muestra que posee una abundancia desusadamente alta de los elementos carbono, nitrógeno y oxígeno. Presumiblemente, esos elementos existían en las capas interiores de la enana blanca justo antes de la explosión.

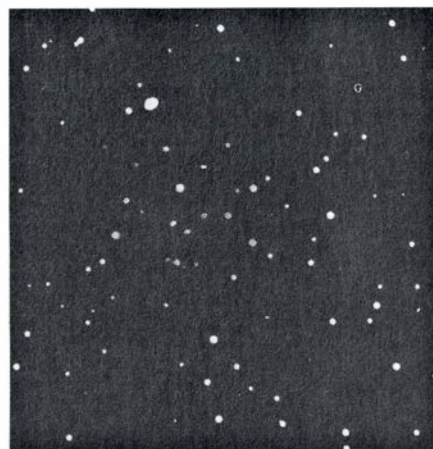
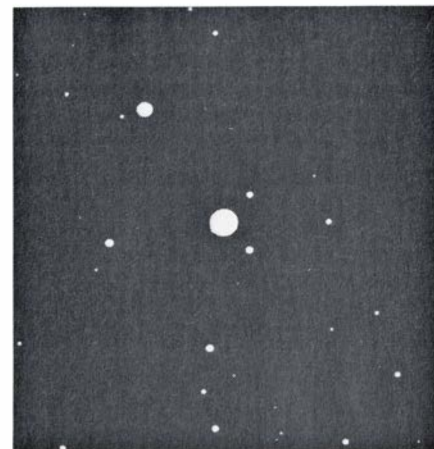
toriamente, cada uno en órbita alrededor del otro, se puede calcular a partir de sus masas y de su período orbital. Las masas de la mayoría de las estrellas caen dentro de un margen bastante estrecho, desde una décima parte hasta 10 veces la masa del Sol; así pues la distancia entre los dos cuerpos en el sistema DQ Herculis podía estimarse en forma bastante aproximada, aun sin conocer específicamente las masas de ambas estrellas. Habida cuenta del período del sistema, muy corto, resultaba evidente que las dos estrellas debían estar muy cercanas entre sí, separadas por algo más del diámetro de una estrella típica.

Los astrónomos hallaron este resultado apasionante. ¿Era sólo una coincidencia el que la nova de 1934 fuese miembro de un sistema binario muy compacto? ¿O estaba la explosión relacionada, de alguna manera, con la estrecha proximidad de ambas estrellas? Se comenzó una búsqueda para detectar posibles estrellas compañeras en otros sistemas de nova conocidos. Todas las novas conocidas están demasiado lejanas para que un sistema doble muy próximo se pueda resolver en sus estrellas individuales y, por tanto, hubieron de emplearse métodos indirectos de detección. La existencia de sistemas binarios se puede deducir de dos maneras: porque una estrella eclipsa a la otra, como en el caso de DQ Herculis, o por la observación de un corrimiento Doppler en las líneas espectrales de una estrella en tal sistema, resultante del movimiento de la estrella alrededor de su compañera.

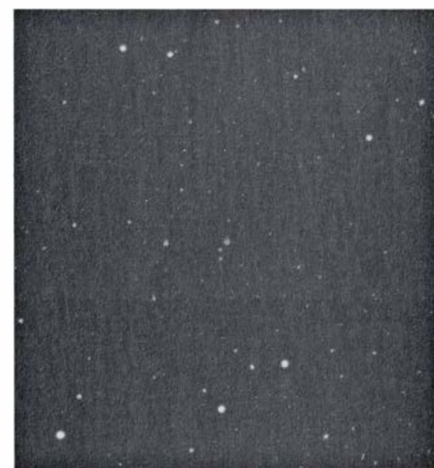
Robert P. Kraft, de los Observatorios de M. Wilson y M. Palomar, estudió 10 antiguas novas y halló que siete presentaban claras indicaciones de la presencia de una estrella compañera de la que no se tenía noticia. En la mayoría de los casos en que se pudieron determinar los períodos orbitales de los sistemas binarios, resultaron inferiores a un día. Las observaciones de Kraft no podían revelar un sistema binario si el plano de las órbitas del sistema era casi perpendicular a la visual del observador, pero parecía razonable suponer que todas las antiguas novas que Kraft estudió eran miembros de sistemas binarios y que las distancias que separaban las dos estrellas eran, generalmente, comparables al diámetro de una estrella normal. En cualquier caso, sus observaciones constituían poderosas indicaciones circunstanciales de que el fenómeno de la nova estaba directamente relacionado

NOVA DQ HERCULIS (*ilustración de la derecha*), **fotografiada en 1934 en el Observatorio Yerkes, de la Universidad de Chicago, cuando su magnitud visual alcanzó 1,4. Antes de la explosión** (*centro de la fotografía de la izquierda*), **su magnitud era 14. La estrella aumentó pues su brillo 100.000 veces.**

NOVA AQUILAE DE 1918, la más brillante nova observada en los últimos 100 años, alcanzó una magnitud visual de −1,1, lo que la convirtió, por unos días, en la segunda estrella del cielo en cuanto a brillo. En la fotografía de la izquierda, que se tomó antes de la explosión, la futura nova es la débil estrella de magnitud 10,6 que aparece en la parte central de la fotografía. La fotografía de la derecha muestra Nova Aquilae cerca de su máximo brillo. Durante su explosión, la nova aumentó su brillo en unas 12 magnitudes, es decir, se multiplicó por 60.000. Las fotografías se tomaron en el Observatorio de Yerkes.

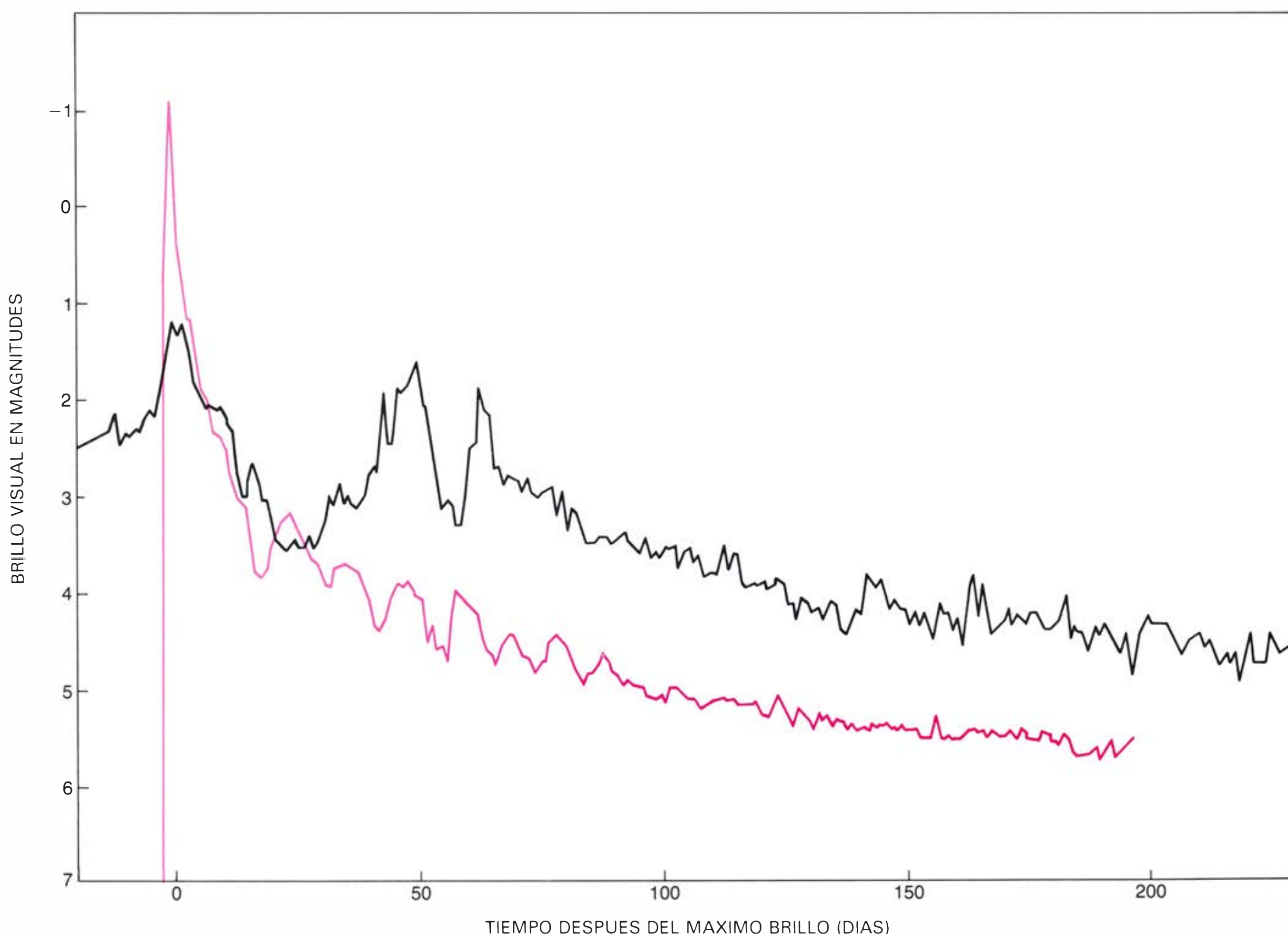
NOVA CYGNI de 1975, la nova brillante más reciente, fotografiada (*izquierda*), **en el Observatorio Lick, el 31 de agosto de 1975, cuando había alcanzado su magnitud máxima de 1,8. En la fotografía de la derecha, tomada tres meses más tarde, la nova se había debilitado hasta la magnitud 11, es decir, se había dividido por 4800 respecto al máximo. El brillo de la estrella antes de la explosión es incierto, porque la estrella era demasiado débil para poder registrarla fotográficamente. Por consiguiente, su brillo hubo de ser inferior al de las estrellas de magnitud 20; su aumento de brillo debe haber correspondido, pues, a un factor de, por lo menos, 19 millones, lo que constituye una auténtica plusmarca para explosiones de nova.**

con la existencia de sistemas de estrellas dobles muy próximas.

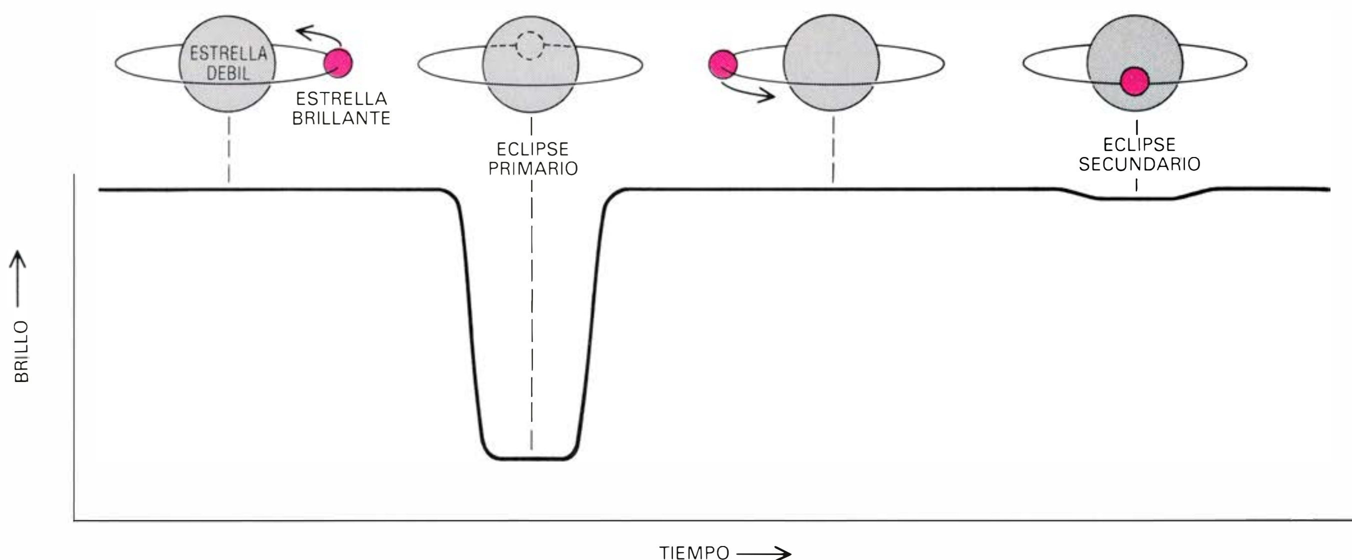
En aquellos casos en que las observaciones de novas antiguas revelaron un sistema eclipsante, se pudo deducir información adicional acerca de la naturaleza de las estrellas que lo constituían. Si se conocen las velocidades orbitales de las estrellas, a través de los corrimientos Doppler de sus líneas espectrales, los tamaños de las estrellas se pueden calcular por medio de la duración de los eclipses. Además, podemos hallar los brillos relativos de las dos estrellas comparando la intensidad del sistema cuando las estrellas se están eclipsando, una a la otra, con su intensidad cuando no se eclipsan. Esta clase de análisis de los sistemas de antiguas novas condujo a otro interesante e inesperado resultado: es corriente que una de las estrellas de un sistema binario de nova sea moderadamente caliente y muy pequeña. De hecho, las compañeras pequeñas eran menores que lo que ninguna estrella normal podía ser. Sólo podía tratarse de enanas blancas extremadamente densas: estrellas cuya masa típica es de alrededor de dos tercios de la del Sol y que tienen aproximadamente el tamaño de la Tierra.

Una enana blanca es una estrella que está al borde del final de su ciclo de evolución. Previamente, ha convertido la casi totalidad de su combustible de hidrógeno y helio en carbono y oxígeno, mediante fusión nuclear, y ya no puede producir energía. Faltándole una fuente de energía para mantenerse a sí misma frente a la atracción gravitatoria mutua de los átomos que la constituyen, se contrae hasta un estado extremadamente denso y compacto, en el que la materia ya no se puede comprimir más. De la materia extremadamente densa de tal estrella, más parecida a un sólido que a un gas normal, se dice que está degenerada. La incompresibilidad de un gas degenerado impide que la estrella se comprima más, llegándose, así al final de la evolución estelar. Normalmente, una enana blanca se enfría lentamente, a lo largo de miles de millones de años y acaba por desaparecer en la oscuridad.

El descubrimiento de que la mayoría de las novas parecen ser miembros de sistemas binarios muy próximos entre sí, en los que uno de los miembros es una enana blanca, sugirió una explicación plausible de las explosiones de nova. De acuerdo con las ideas establecidas por Evry Schatzman, del Instituto de Astrofísica de París, y Leon Mestel, de la Universidad de Leeds, Kraft propuso que las explosiones de nova las causaba por la transferencia de materia de la estrella normal de un par binario sobre la compañera enana blanca degenerada. En razón de su elevada densidad, la enana blanca tiene un campo gravitatorio muy intenso, y el gas que tiende a depositarse en su superficie es acelerado hasta velocidades muy altas. La materia que choca con la superficie de la enana blanca experimenta, por tanto, un calentamiento hasta tempera-



DIFERENCIA entre las novas rápidas y las lentas; puede apreciarse en las curvas de luz de Nova Aquilae 1918 (*en color*) y Nova Pictoris 1925 (*negro*). Ambas novas igualaron por breve tiempo el brillo de las estrellas más brillantes. (Un cambio de cinco magnitudes representa un aumento de 100 veces en brillo.) El incremento de éste en Nova Aquilae, muy rápido, fue seguido de un abrupto descenso durante el cual se observaron fluctuaciones periódicas. En abierto contraste, Nova Pictoris tardó varios meses en alcanzar su máximo brillo, después de lo cual se atenuó en forma más irregular y gradual que Nova Aquilae. Nova Pictoris fue descubierta por un vigilante nocturno en Africa del Sur, que la observó por primera vez cuando iba de regreso a casa.



DEBILITAMIENTO PERIODICO de una estrella brillante. Nos indica que la estrella es miembro de un par binario eclipsante. Esa prueba ha establecido que los sistemas de nova son binarios, donde uno de los miembros es una estrella enana blanca. Para observar el debilitamiento, importa que las estrellas estén orientadas de manera que el plano de sus órbitas se vea casi "de canto". Se apreciará entonces cómo ambas estrellas se eclipsan a intervalos regulares. El brillo relativo del sistema durante los dos eclipses guarda una relación directa con los brillos de las superficies de ambas estrellas. Además, el tiempo que tarda cada eclipse en hacerse total, comparado con la duración del eclipse entero, sirve para determinar los tamaños relativos de ambas estrellas. No hay ningún sistema binario de nova suficientemente próximo al sistema solar para poder observar por separado las dos estrellas miembros.

turas extremadamente altas. Con el tiempo, la temperatura se eleva lo suficiente para desencadenar reacciones nucleares en la superficie de la enana, lo que conduce a una violenta liberación de energía.

Este brusco desprendimiento de energía lo permite el hecho de que las reacciones nucleares en la superficie de la enana blanca ocurran en un gas degenerado. En un gas normal no puede producirse la liberación explosiva de energía porque, al aumentar la temperatura, el gas se dilata. Con la dilatación, decrece la temperatura del gas, y las velocidades de las reacciones nucleares, que son función de la temperatura, disminuyen con ésta. Este efecto termostático permite a las estrellas normales radiar energía a un régimen constante durante miles de millones de años.

La materia degenerada, a causa de su incompresibilidad, no se dilata al calentarse. Por tanto, cuando tienen lugar las reacciones nucleares, la temperatura de la materia aumenta continuamente, lo que da por resultado que las reacciones nucleares se efectúen a mayores velocidades. Esto calienta el gas todavía más, de modo que el proceso crece sin límite, originando series de reacciones nucleares desbocadas, o inestables. La producción de energía crece hasta niveles enormes, culminando en un brote explosivo.

La cesión de gas desde una estrella normal a su compañera próxima, compacta y degenerada, no es exclusiva de las novas. Un proceso semejante da lugar a las estrellas de rayos X, o sistemas binarios, en los que el objeto compacto no es una estrella enana blanca, sino que se trata, según se cree, de una estrella de neutrones o de un agujero negro. La materia en una estrella de neutrones está mucho más condensada (y, en un agujero negro, infinitamente más condensada) que en una enana blanca. Como resultado, el gas que describe espirales en un disco de acumulación de materia hacia el objeto compacto se acelera hasta velocidades extremadamente altas. En virtud de la creciente energía cinética adquirida por el gas, la radiación asociada con el disco se emite primordialmente en la región de rayos X del espectro y no en la región óptica.

Gran parte del trabajo teórico sobre las novas, realizado en los últimos diez años se ha dedicado a comprobar si una explosión de nova constituye un proceso termonuclear sin control en la superficie de una enana blanca de un sistema binario estrechamente ligado. Los principales defensores de la hipótesis han sido Sumner G. Starrfield, de la Universidad del estado de Arizona, James W. Truran, Jr., de la Universidad de Illinois, en Urbana-Champaign, y Warren Sparks, del Centro de Vuelos Espaciales Goddard. Sus extensos cálculos del fenómeno de la nova concuerdan, en general, con cierto número de importantes características de las explosiones. El panorama que surge del trabajo que ellos, y otros, han desarrollado se puede resumir de la manera siguiente.

El depósito de gas sobre la enana blanca es continuo, impulsado por la constante dilatación de las envolturas externas de la estrella compañera, a medida que ésta evoluciona hacia su conversión en una gigante roja. El gas, que se acumula en la forma de un disco en el plano de la órbita de la estrella, está compuesto principalmente por hidrógeno. Cuando cae en espiral a alta velocidad sobre la superficie de la enana blanca, se mezcla con la materia que constituye las capas exteriores de la enana: principalmente helio, quizá con alguna mezcla de carbono y oxígeno, que se han originado en el interior de la enana.

Normalmente, las enanas blancas carecen de hidrógeno porque se ha convertido en los elementos más pesados, helio, carbono y oxígeno, mediante reacciones nucleares en las etapas iniciales de la evolución de la estrella. La creación de una nueva provisión de hidrógeno sobre la superficie de la enana es importante porque éste, al poseer un núcleo muy sencillo (un protón), es el elemento que reacciona más fácilmente con otros núcleos para liberar energía. La adición de nuevo hidrógeno a los elementos más pesados, ya presentes en la enana, posibilita una nueva serie de reacciones nucleares. La iniciación de las reacciones de fusión requiere una temperatura de unos 20 millones de grados Kelvin. A

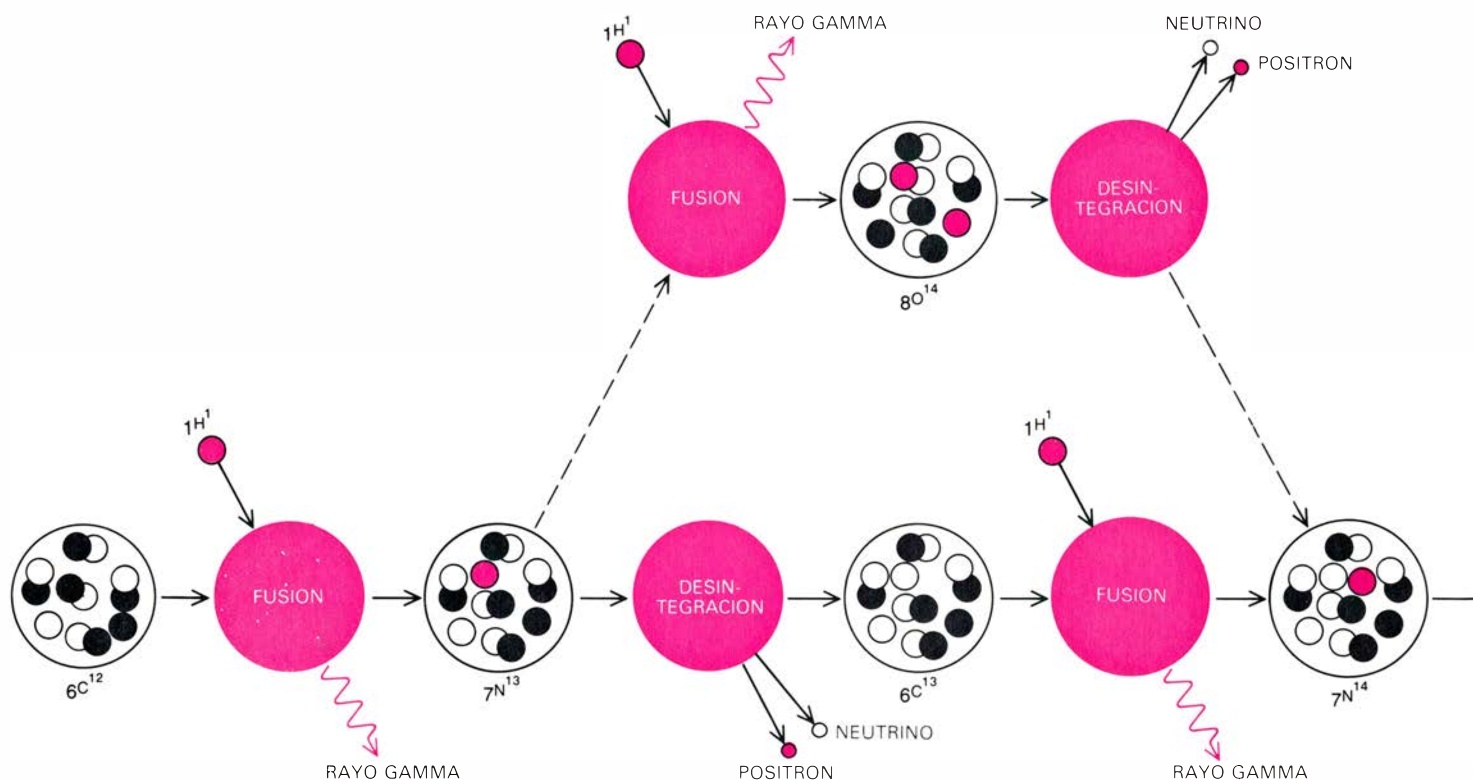
esa temperatura, los distintos núcleos son acelerados hasta velocidades suficientemente altas para superar la repulsión electrostática ejercida por sus cargas positivas, posibilitando que el núcleo choque, se funda y desprenda energía.

La temperatura necesaria se logra gracias al impacto del gas que cae, en espiral y a gran velocidad, sobre la superficie de la enana blanca. Aunque el tiempo que tarda la temperatura en alcanzar el nivel requerido para la fusión depende de varios factores, incluyendo la velocidad a la cual la estrella compañera en expansión suministra gas al disco de acumulación, es típicamente del orden de decenas de miles de años. Con el tiempo, se alcanza una temperatura a la cual la materia degenerada en la superficie de la enana blanca, sembrada con nuevo hidrógeno, empieza a mantener reacciones nucleares.

En las reacciones de fusión, los núcleos de los elementos más pesados capturan protones. Los núcleos recién fundidos se desintegran luego a velocidades fijadas por sus características vidas medias radiactivas. En las reacciones postuladas para la superficie de las enanas blancas, la desintegración suele dar por resultado la liberación de un electrón positivo (positrón) y un neutrino. La cadena de reacciones es exactamente la misma que la que convierte hidrógeno en helio en el interior de estrellas normales de mayor masa que el Sol. La sucesión de choques, que aboca en la conversión neta de cuatro núcleos de hidrógeno en uno de helio, afecta a los núcleos de los elementos carbono, nitrógeno y oxígeno como catalizadores; por cuya razón se le denomina ciclo CNO [*véase la ilustración inferior*].

Los cálculos de Starrfield, Truran y Sparks han demostrado que las reacciones termonucleares del ciclo CNO comenzarán en la superficie de una enana blanca que acumule suficiente hidrógeno a lo largo de un período de tiempo determinado. Los cálculos muestran que, dadas las condiciones que se espera prevalezcan en sistemas binarios estrechamente ligados, el comienzo de las reacciones nucleares conduce a una rápida reacción termonuclear no controlada en la materia degenerada de la superficie. Algunas características de esos modelos teóricos concuerdan bien con los caracteres observados en las novas, tales como la forma de las curvas de luz y la energía total generada. Los cálculos también indican que una de las variables más importantes en la determinación de las características de la explosión de nova es la abundancia de carbono, nitrógeno y oxígeno con respecto la concentración de hidrógeno. Una materia degenerada, rica en esos elementos más pesados, genera energía mucho más rápidamente y en mayores cantidades que una materia apreciablemente menos rica en ellos. De hecho, los modelos señalan que una de las importantes diferencias entre novas rápidas y lentas radica en la composición química de la materia de la superficie de la enana blanca.

De acuerdo con los estudios teóricos, si el carbono, nitrógeno y oxígeno son aproximadamente 100 veces más abundantes con respecto al hidrógeno de lo que son en las estrellas normales semejantes al Sol (es decir, si el número de átomos de carbono, nitrógeno y oxígeno juntos es igual al 1 por ciento del número de átomos de hidrógeno), las reacciones nucleares que se registran en el gas degenerado ocurren a tal velocidad que la temperatura en la superficie alcanza rápidamente los 100 millones de grados K y, en el plazo de un minuto, la energía se libera explosivamente. La erupción, en forma de radiación y de partículas, abarca la superficie de la enana blanca. En el intervalo de unas horas, una importante proporción de la capa exterior depositada sobre la enana es expulsada



CADENA DE REACCIONES NUCLEARES, responsable de las explosiones de nova, convierte hidrógeno en helio por el mismo proceso que ocurre en el núcleo de las estrellas normales. Cuatro núcleos de hidrógeno, o protones, se funden, uno tras otro, con núcleos de carbono, nitrógeno y oxígeno en la sucesión de reacciones conocida por ciclo CNO. Las tres primeras reacciones de fusión en el ciclo van acompañadas de emisión de rayos gamma de gran energía. La última reacción de fusión, en la que un protón reacciona con un núcleo de nitrógeno 15, va seguida de la expulsión de un núcleo de helio, o partícula alfa ($_2He^4$), y la regeneración de un núcleo de carbono 12, para iniciar nuevamente el ciclo. Parte de las reacciones de captura de protones

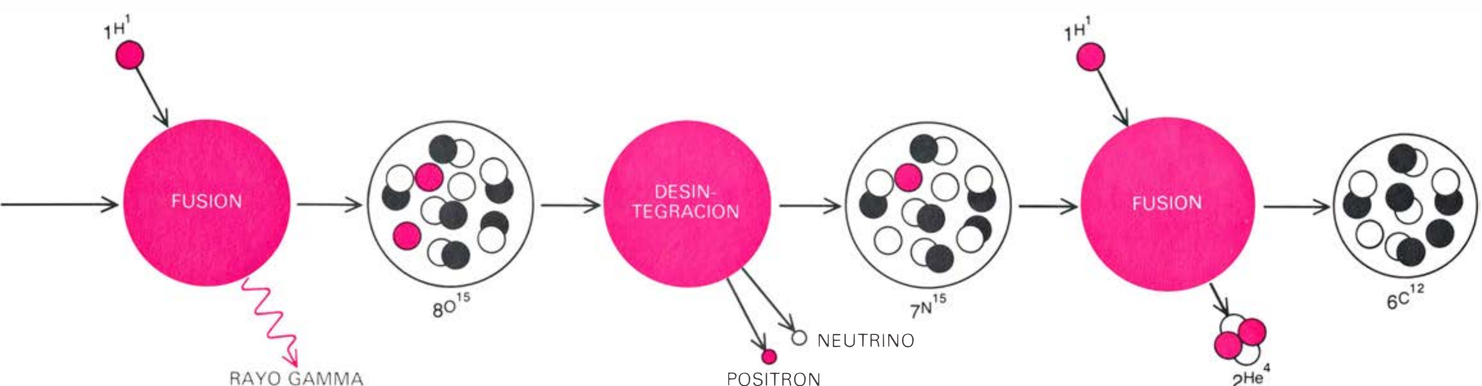
hacia el espacio y la estrella aumenta de brillo con gran celeridad, de una manera muy semejante a como lo hace una nova rápida. Por otro lado, los modelos indican que, si las capas superficiales degeneradas de la enana blanca tienen concentraciones de carbono, nitrógeno y oxígeno no superiores a las que se encuentran en las estrellas normales, la reacción termonuclear no controlada se extiende a un intervalo de tiempo más largo y la nova es de carácter lento.

La predicción de que las diferencias en composición química puedan ser, en gran parte, responsables de los diferentes tipos de explosiones de nova proporciona una oportunidad de comprobar la validez de los modelos y el cuadro básico de la explosión. Quizás, el método más directo de establecer la composición del gas en la superficie de una enana blanca sea estudiar las envolturas gaseosas que son expulsadas en una explosión. Una vez que tales envolturas han sido expulsadas, no las perturba actividad ulterior alguna relacionada con el sistema de la nova; por tanto, deben reflejar las condiciones que condujeron a su formación.

Los espectros de las novas durante una explosión siempre presentan indicaciones de que la materia ha sido expulsada rápidamente del sistema de nova. Las envolturas expulsadas están compuestas de una proporción apreciable de materia que había sido depositada sobre la enana blanca procedente de su compañera mayor, llegando a un 0,01 por ciento de la masa total de la enana. La materia expulsada, disparada hacia el espacio a velocidades de 1000 kilómetros por segundo, forma una envoltura en expansión alrededor del sistema binario. Tales envolturas se disipan rápidamente, a medida que se expanden, desapareciendo generalmente de la observación visual antes de adquirir un tamaño que permita su detección. Sin embargo, se han observado las capas expulsadas por algunas de las novas más próximas a nosotros.

En el Observatorio Steward, de la Universidad de Arizona, mis colaboradores y yo hemos emprendido un programa para buscar las envolturas de antiguas novas y examinarlas. Los modelos han indicado que la materia expulsada de las novas rápidas debe ser rica en carbono, nitrógeno y oxígeno, y que las capas que rodean las novas lentas no deben poseer tal riqueza. La composición de una envoltura y las condiciones físicas en que se halla se pueden determinar examinando su espectro. Un gas enrarecido, tal como la envoltura de una nova, radia una línea brillante, o espectro de emisión. El gas radia solamente a ciertas longitudes de onda bien definidas, determinadas por su composición, densidad y temperatura. Midiendo las intensidades relativas de las líneas de emisión a varias longitudes de onda, se puede obtener información acerca de las abundancias de los elementos y la temperatura de la envoltura.

La primera de tales capas que analizamos espectroscópicamente fue la envoltura simétrica alrededor de DQ Herculis, la nova lenta que apareció en 1934. Encontramos que el espectro de la nova, obtenido con el telescopio de 2,3 metros del Observatorio Steward, no se parecía a ningún otro conocido hasta entonces. Por regla general, los espectros de las nubes de gas radiante en el espacio, tales como la Gran Nebulosa de Orión, son muy parecidos entre sí. Las nebulosas brillan a causa de la energía que reciben de las estrellas calientes, con las que están asociadas. Las estrellas calientes radian primordialmente en la región ultravioleta del espectro, y cuando esta radiación de gran energía ilumina el gas alrededor de la estrella, lo ioniza, es decir, arranca electrones de los átomos del gas. Por tanto, el gas se calienta hasta unos 10.000 grados Kelvin, y los choques entre los iones y los electrones li-



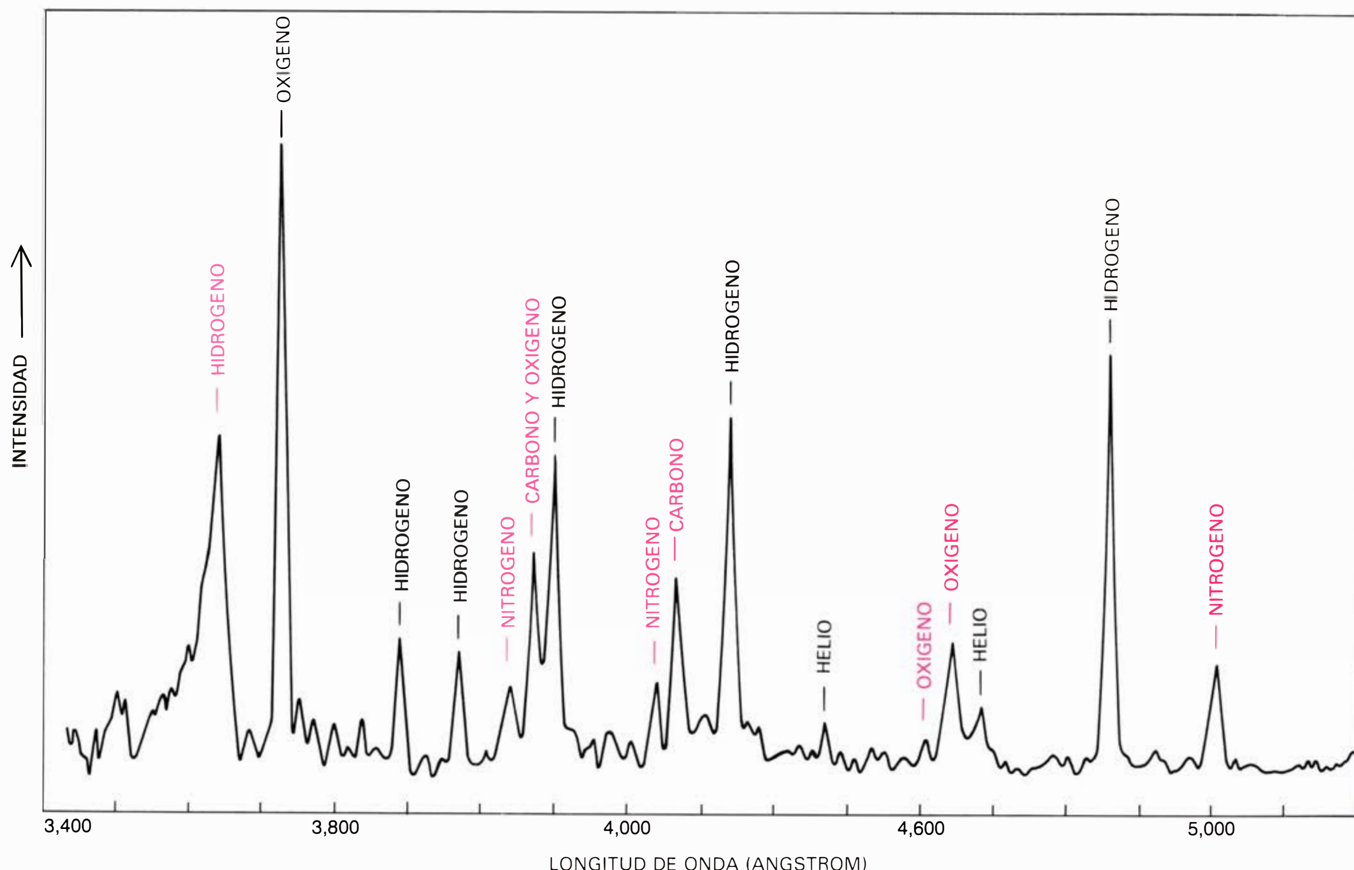
van seguidas de reacciones de desintegración que también desprenden energía en la forma de un positrón (electrón positivo) de mucha energía y de un neutrino. (La emisión de un positrón transforma un protón nuclear en un neutrón.) La velocidad de las reacciones de fusión depende de la temperatura del gas. Una vez que las reacciones comienzan en el gas degenerado de la superficie de una enana blanca, aumentan la temperatura del gas, lo que eleva la velocidad de las reacciones. Con el tiempo, el resultado es el desencadenamiento de una reacción nuclear incontrolada que da lugar a la explosión. El que la explosión resulte lenta o rápida depende de cuáles sean las concentraciones de carbono, nitrógeno y oxígeno en el gas degenerado

berados producen la radiación del espectro de emisión característico de la nebulosa.

El espectro de la envoltura de DQ Herculis era completamente especial: muchas de las más intensas líneas halladas en él no tenían contrapartida en los espectros de emisión de las nebulosas. El misterio del origen de la emisión procedente de la envoltura únicamente se resolvió cuando nos dimos cuenta de que uno de los rasgos del espectro, originado por el hidrógeno, requería que el gas estuviera mucho más frío que el gas de las nebulosas de emisión, es decir, a unos 500 grados Kelvin. Esto explicaba por qué muchas de las líneas espectrales que resultan familiares en las nebulosas no aparecían en el espectro de la envoltura. Cuando finalmente se identificaron los átomos emisores, resultaron ser iones de hidrógeno, helio, carbono, nitrógeno y oxígeno. También aquí surgió la sorpresa: para ionizar los átomos se necesitan altas energías y éstas van asociadas, generalmente, a altas temperaturas.

La paradójica ionización de un gas frío tiene explicación en el caso de que la envoltura de la nova no esté en equilibrio, sino recuperándose todavía de los efectos de la explosión. Se sabe, a partir de observaciones realizadas en tiempos próximos a la explosión de una nova, que la capa expulsada está altamente ionizada y caliente, con temperaturas que exceden los 15.000 grados Kelvin. Con el paso del tiempo, la superficie de la enana blanca se enfría y deja de emitir la intensa radiación ultravioleta que originariamente calentó e ionizó la envoltura. En algunas ocasiones, hay otras fuentes de energía a disposición de la envoltura en expansión. Si ésta se está dilatando en una región en la que el tenue gas del espacio interestelar es relativamente denso, el choque entre ella y el gas interestelar puede ser de gran energía. Tal choque puede muy bien explicar la infrecuente apariencia de la envoltura de GK Persei, una nova brillante de 1901, que dio lugar a una de las envolturas más extensas y calientes que se conocen. La envoltura de DQ Herculis de 1934, por otro lado, se está expandiendo sin interferencia por parte del gas interestelar; se extiende fuera del plano central de nuestra galaxia, donde se halla la mayor parte del gas interestelar. Libre para dilatarse sin ningún suministro adicional de energía, la envoltura de DQ Herculis se ha estado enfriando continuamente. Los electrones que fueron arrancados de los átomos los están recapturando ahora los iones. El tiempo necesario para que los electrones sean completamente recapturados es, sin embargo, substancialmente más largo que el necesario para que la temperatura del gas descienda. Resulta así que la envoltura se ha enfriado hasta una temperatura mucho más baja que la suya original y, sin embargo, no todos los electrones han sido recapturados; de aquí la aparente paradoja de un gas frío que está, aún, altamente ionizado.

Al cerrar nuestro examen del espectro de la envoltura, vimos que las abundancias combinadas del carbono, nitrógeno y oxígeno, con respecto al hidrógeno, eran unas cien veces más altas que en las estrellas normales. De acuerdo con los modelos corrientes de las novas, tales abundancias deberían



ESPECTRO DE LA ENVOLTURA DE DQ HERCULIS, recientemente obtenido por el autor y sus colaboradores con el telescopio de 2,3 metros, del Observatorio Steward; no se parece al de ninguna nebulosidad conocida. El espectro de la envoltura de esta nova lenta consta de un débil fondo continuo de radiación, sobre el cual están superpuestas líneas emitidas por elementos en diferentes estados de excitación. La radiación procedente de los elementos rotulados en negro es, generalmente, dominante en los espectros de las nebulosas normales, mientras que las líneas de emisión rotuladas en color suelen ser débiles o inexistentes. La intensidad de la emisión del carbono, nitrógeno y oxígeno muestra que la envoltura de DQ Herculis es desusadamente rica en esos elementos. La intensa línea de emisión del hidrógeno a 3645 unidades Ångström requiere que el gas esté muy frío, a unos 500 grados Kelvin. La emisión del oxígeno a 3727 Ångström constituye una anomalía; está generalmente asociada con gas, cuya temperatura es de 10.000 grados Kelvin.

NOVA	AÑO DE LA EXPLOSION	MAGNITUD MAXIMA	CLASE	TAMAÑO ANGULAR DE LA ENVOLTURA (SEGUNDOS DE ARCO)	TEMPERATURA DE LA ENVOLTURA (GRADOS KELVIN)	ELEMENTOS ENRIQUECIDOS EN LA ENVOLTURA
T AURIGAE	1891	4,2	LENTA	20	≤3.000	HELIO, NITROGENO Y OXIGENO
GK PERSEI	1901	0,2	RAPIDA	75	>25.000	NITROGENO
V476 CYGNI	1920	2,0	RAPIDA	10		
RR PICTORIS	1925	1,2	LENTA	25	15.000	HELIO Y NITROGENO, POSIBLEMENTE NEON
DQ HERCULIS	1934	1,4	LENTA	20	500	CARBONO, NITROGENO Y OXIGENO
CP PUPPIS	1942	0,2	RAPIDA	15	≤1.000	NITROGENO
V533 HERCULIS	1963	3,0	RAPIDA	5		
T PYXIDIS	1890 1944 1902 1966 1920	6,6	LENTA	10	~10.000	NINGUNO

OCHO NOVAS CON ENVOLTURAS OBSERVABLES sujetas a estudio. Aunque hay indicaciones de que todas las novas expulsan gas durante la explosión, son pocas las que producen envolturas detectables. La mayoría de los productos gaseosos expulsados resultan demasiado débiles para aparecer en las fotografías, porque las envolturas tienden a desaparecer antes de alcanzar tamaño observable. Las envolturas se forman alrededor de las novas rápidas y de las lentas; muestran una amplia variación de temperaturas. Salvo la nova recurrente T Pyxidis, las novas parecen crear envolturas con un rasgo común: las envolturas están enriquecidas en uno o más de los elementos más pesados, en comparación con la composición de las estrellas normales.

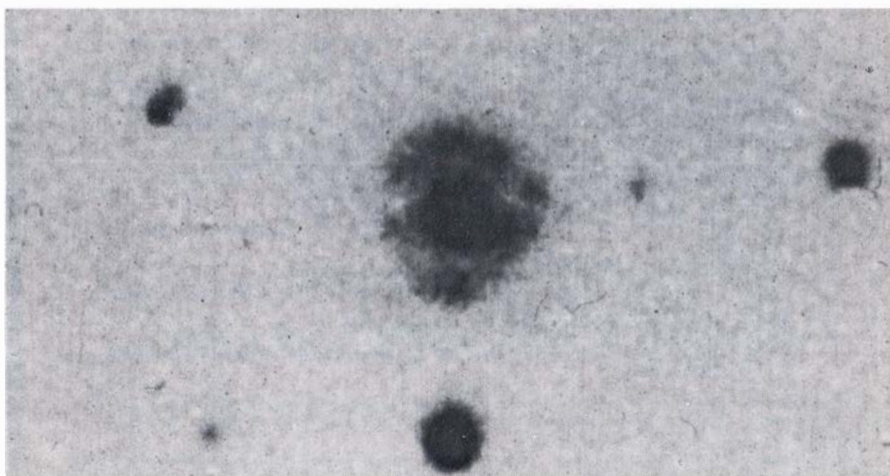
corresponder a las de una nova rápida. Por otro lado, como DQ Herculis es una nova lenta, el análisis espectral y los modelos parecen contradecirse.

Se debe andar con cuidado a la hora de generalizar los resultados de un único objeto a otras novas. John S. Gallagher, de la Universidad de Illinois en Urbana-Champaign, y yo estamos examinando ahora otros sistemas de nova que presentan envolturas extensas. Hemos finalizado el trabajo referente a otros dos objetos adicionales, las envolturas que rodean a las novas lentas RR Pictoris, que explotó en 1925, y T Aurigae, que lo hizo en 1891. Los espectros muestran que la envoltura de RR Pictoris está notablemente enriquecida en helio y nitrógeno; la envoltura de T Aurigae lo está en helio, nitrógeno y oxígeno. Gary J. Ferland y Gregory A. Shields, de la Universidad de Texas en Austin, han llevado a cabo un estudio similar de V1500 Cygni, la nova rápida de 1975, y han encontrado que su envoltura, todavía demasiado pequeña para poderse resolver en las fotografías, es mucho más rica que las estrellas normales en no pocos de los elementos más pesados (carbono, nitrógeno y oxígeno, entre otros).

Hay varias envolturas más de nova accesibles al análisis, algunas de ellas en el cielo del Hemisferio Sur, cuya composición está en estudio. Sobre la base de los resultados ya obtenidos, parece que la mayoría de las novas, tanto lentas como rápidas, se muestran enriquecidas en varios de los elementos más pesados. Aunque se desconoce la fuente del enriquecimiento, es probable que se trate del interior de las enanas blancas, que se sabe es rico en carbono y oxígeno.

No hay que sorprenderse, en absoluto, de la abundancia de elementos pesados en las novas rápidas. Pero el descubrimiento de que las envolturas de las novas lentas DQ Herculis, RR Pictoris y T Aurigae estén enriquecidas en elementos pesados, no concuerda con los cálculos teóricos. En particular, la envoltura de DQ Herculis contiene cantidades de carbono, oxígeno y nitrógeno, tales que, si hubieran existido antes de la explosión, tendrían que haber suministrado suficiente energía para hacer de DQ Herculis una nova rápida.

La dificultad de reconciliar los enriquecimientos observados en las envolturas de novas lentas con las predicciones del modelo CNO ha inducido a algunos teóricos a especular sobre la posibilidad de que las erupciones de nova no siempre se ajusten al ciclo CNO de reacciones. Podrían explicarse los enriquecimientos si las erupciones de nova fuesen generadas por reacciones nucleares capaces de sintetizar carbono, nitrógeno, oxígeno y otros elementos directamente, a partir del hidrógeno y el helio, cosa que no hace el ciclo CNO. Hay, por ejemplo, reacciones termonucleares en las que tres núcleos de helio (partículas alfa) se funden para dar carbono 12, del cual se pueden formar otros elementos más pesados por subsiguiente captura de protones. Este



LA NOVA RECURRENTE T PYXIDIS ha sufrido cinco explosiones desde 1890, la más reciente de las cuales ocurrió en 1966. Las novas recurrentes multiplican su brillo por un factor de sólo 1000 durante una explosión, aumento harto menor que el de una nova normal. Se acaba de descubrir que las novas recurrentes también expulsan envolturas observables (como se evidencia en este negativo de una fotografía de T Pyxidis, obtenida electrónicamente con un tubo de formación de imágenes en el telescopio de cuatro metros del Observatorio de Kitt Peak). La envoltura fue descubierta en 1978 por Harvey R. Butcher, de Kitt Peak, D. A. Kopriva y el autor, en una búsqueda de envolturas alrededor de las novas.

proceso "triple alfa" es, como se sabe, una importante fuente de energía en las estrellas gigantes rojas. Probablemente, la superficie degenerada de las enanas blancas contiene helio suficiente para que se produzca esta reacción; parece posible, pues, que algunas explosiones de nova fueran debidas al proceso triple alfa más que al ciclo CNO, basado en el hidrógeno.

La abundancia de helio en los discos de acumulación antes de la explosión de nova, comparada con su abundancia en las envolturas después de la explosión, puede suministrar indicaciones adicionales sobre la fusión del helio en algunas explosiones de nova. Estudios recientes de novas antiguas acometidos en el Observatorio Steward han demostrado que la cantidad de helio en las envolturas expulsadas es inferior a la que se encuentra comúnmente en el gas que se está depositando sobre las enanas blancas. Una explicación lógica para esta aparente disminución en el contenido de helio del gas es que se hayan producido reacciones triple alfa durante la explosión, convirtiendo helio en carbono y otros elementos.

A pesar de las incertidumbres sobre algunos detalles, existe ahora acuerdo general en que las erupciones de nova son probablemente producidas por reacciones termonucleares en la superficie de las enanas blancas en sistemas binarios estrechamente ligados. El proceso se inicia con la cesión de materia desde una estrella compañera en expansión. La materia, en su caída, incide sobre la superficie degenerada de la enana blanca a una velocidad tan alta que la superficie se calienta hasta la temperatura de 20 millones de grados, necesaria para desencadenar reacciones nucleares incontroladas. Los modelos de nova basados en las reacciones del ciclo CNO han dado cuenta, con éxito, de muchas de las características observadas de las novas. Sin embargo, recientemente se han obtenido nuevos datos en torno a la composición de envolturas de novas antiguas que requieren modificar algunas de las anteriores ideas sobre la naturaleza de las explosiones. Una investigación ulterior en el ámbito de antiguas envolturas de novas resultará estimulante, no sólo por la información que, inevitablemente, proporcionará de las novas, sino también porque las envolturas han conducido ya a los astrónomos a un medio circundante poco corriente y completamente distinto de todos los encontrados hasta ahora en nuestra galaxia.

Insectos filtradores

Insectos pertenecientes a tres órdenes hacen eclosión bajo el agua y capturan el alimento con redes, pinceles y otras estructuras. Desempeñan un importante papel al oponerse a la tendencia de los ecosistemas a perder materia orgánica

Richard W. Merritt y J. Bruce Wallace

Muchos insectos, en especial la larva de polilla a la que llamamos gusano de seda, hilan filamentos para construir un capullo. Menos conocidos son los insectos que hilan filamentos para atrapar alimento. No sólo tejen redes de fina malla para recolectar la materia orgánica que ingieren, sino que, además, realizan toda la tarea bajo el agua. Los hiladores de redes pertenecen al grupo de insectos filtradores que hacen eclosión en ríos, lagos y otros ambientes acuáticos y pasan su vida inmatura completamente sumergidos. De los 27 órdenes de insectos, diez tienen representantes acuáticos, pero sólo tres incluyen especies que se alimentan realmente por filtración: las moscas verdaderas (del orden Dípteros), las frigáneas (del orden Tricópteros) y las efímeras (del orden Efemerópteros).

Las moscas verdaderas y las frigáneas son endopterigotos: carecen de alas hasta el estadio de pupa, momento en el que se desarrollan las estructuras alares, al tiempo que la larva asume su forma adulta. Las efímeras son exopterigotos: las estructuras alares están presentes desde el instante en que el organismo sale del huevo hasta que alcanza la edad adulta, la apariencia del insecto inmaturo augura, en general, su morfología adulta y éste no pupa en absoluto. Las larvas de los endopterigotos mudan varias veces antes de llegar al estadio pupal. Los juveniles exopterigotos, a los que generalmente se denomina ninfas, aumentan asimismo de tamaño a través de una serie de mudas; el insecto adulto emerge después del estadio inmaturo final.

Sea cual fuere su ciclo biológico, varias especies de moscas, frigáneas y efímeras ocupan hábitats acuáticos tan diversos como los rápidos arroyos alpinos, los ríos serpenteantes, los tranquilos fondos de lagos y los estuarios de marea; sus hábitats, con frecuencia, se superponen. Independientemente del orden al que pertenezcan se les suele dividir en dos grupos: el de las especies que viven allí donde las corrientes de agua activas permiten un modo de alimentación pasivo, y el de las que viven en lugares donde las corrientes son mínimas y el propio insecto ha de alterar la quietud del medio.

¿Qué es lo que sirve de alimento a un insecto filtrador? Numerosos análisis de su contenido intestinal indican que no suelen distinguir entre material inorgánico y orgánico hallado en el agua, por lo que ingieren partículas de arcilla y pequeños granos de arena junto con alimentos de origen vegetal (bacterias y algas) y de origen animal (protozoos y pequeños invertebrados). Con mucho, sin embargo, la mayor parte de su dieta está constituida por partículas orgánicas, a menudo de origen no identificable, que colectivamente se conocen con el nombre de detritos finos. Entre las fuentes de las partículas orgánicas se hallan, primeramente, las heces de los insectos acuáticos carroñeros, los "trituradores" que se alimentan de vegetación en descomposición; en segundo lugar, las heces de otros animales acuáticos que hacen presa en animales más pequeños o comen tejido vegetal vivo; tercero, la materia orgánica transportada de la tierra al agua por la escorrentía y, en cuarto lugar, la agregación de materia orgánica que ha dejado de estar en solución. Cada partícula detrítica puede sostener, asimismo, un recubrimiento formado por la flora propia de la descomposición: bacterias, hongos y otros microorganismos.

Los detritos finos, por lo general el alimento más abundante de que disponen los animales filtradores, es el menos remunerador en términos de eficiencia de asimilación, es decir, el porcentaje de alimento ingerido que el animal absorbe. La eficiencia de asimilación de los detritos oscila entre el 2 y el 20 por ciento, frente al 30 por ciento en el caso de algas y más del 70 por ciento si se trata de tejidos de origen animal.

Esta relación de las estrategias de alimentación de los insectos filtradores dará comienzo con aquéllos cuyos hábitats se encuentran en corrientes rápidas. Uno de los mecanismos de filtración menos complejo es el de la ninfa de *Isonychia,* un género de efímeras. Las patas anteriores del insecto tienen un denso fleco de sedas, largas estructuras en forma de cerdas. Cada cerda porta dos hileras de finos pelos. Los pelos de una de las hileras son moderadamente largos, y, cortos y ganchudos, los de la otra. Cuando los pelos ganchudos de una seda se traban con los largos de la siguiente, el filtro formado por los pelos entrelazados puede atrapar partículas de un diámetro inferior a un micrometro (una milésima de milímetro). Para recolectar su alimento, la ninfa de efímera se sujeta a alguna superficie adecuada, se encara a la corriente y levanta las patas anteriores con sus series de sedas entrelazadas. Transcurrido cierto tiempo, el insecto acerca las patas anteriores a las partes bucales, recoge las partículas capturadas y las ingiere.

Las larvas de dos géneros de tricópteros, *Brachycentrus* y *Oligoplectrum,* tienen un sistema de pareja simplicidad. Construyen refugios transportables oblongos trabajando con materiales inorgánicos y restos vegetales, que unen y fijan mediante sus secreciones siriciformes. El extremo abierto del refugio está dirigido corriente arriba. La larva se guarece en el interior del refugio y, cuando está filtrando, extiende a la vez las seis patas fuera de la abertura, en abanico. Sus patas medias y posteriores portan una hilera de cerdas; mientras las partículas alimenticias son capturadas por estos cuatro filtros, el insecto maneja sus patas anteriores pa-

ra peinar las sedas, limpiarlas y formar con las partículas recogidas una pella adecuada para la ingestión. Las larvas de *Brachycentrus* no fían exclusivamente en la filtración; también pastan vegetales microscópicos.

Las larvas de la familia Simulíidos, las moscas negras, han desarrollado un sistema de alimentación que, desde el punto de vista estructural y el etológico, está bien adaptado a un hábitat de corriente rápida. Aunque ápodas, tienen un anillo de ganchos al final de su abdomen. Mediante la ayuda de secreciones siriciformes procedentes de sus glándulas salivales fijan los ganchos a rocas o a plantas sumergidas y después giran el cuerpo de manera que la parte inferior de su cabeza, con sus partes bucales, se dirige a la corriente. Esta postura preferente constituye la manifestación etológica de la adaptación de los insectos a su hábitat.

Un aspecto estructural de la adaptación de la larva de los simulíidos consiste en un par de piezas bucales insólitas denominadas abanicos cefálicos, órganos retráctiles situados entre las antenas y las mandíbulas. Cuando la larva ha adoptado su posición con respecto a la corriente, los abanicos cefálicos se extienden para formar un aparato filtrador considerablemente mayor en superficie que la propia cabeza. Después, los abanicos cefálicos son retraídos uno tras otro y la larva rebaña e ingiere las partículas capturadas con sus mandíbulas.

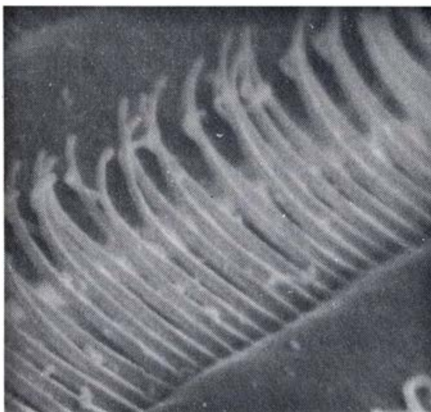
Esta no es la única adaptación estructural de la larva de las moscas negras. El análisis del contenido del intestino larval muestra que capturan algas, partículas detríticas, sus bacterias asociadas y fragmentos de arena y fango en una gama de tamaños que va desde los 350 micrometros hasta los 0,01 micrometros, lo cual es considerablemente menor que la malla del filtro del abanico cefálico. Douglas H. Ross, de la Universidad de Georgia, y Douglas A. M. Craig, de la Universidad de Alberta, han señalado recientemente un posible mecanismo por el que dicho filtro puede capturar las partículas más pequeñas. Resulta que las larvas de simulíidos segregan mucus de unas glándulas situadas en la parte frontal de su cabeza y que el movimiento de sus mandíbulas esparce este material pegajoso sobre la superficie de los abanicos cefálicos. Cuando las pequeñas partículas tocan el mucus al pasar a través del abanico quedan adheridas a él. Se sabe que varios invertebrados filtradores marinos segregan mucus, pero hasta ahora se desconocían estas secreciones entre los insectos acuáticos.

Esto nos lleva a los filtradores men-





ABANICOS GEMELOS en la cabeza de una larva de simulíido según se ven en estas micrografías electrónicas de barrido. A la izquierda, se hallan en su posición retraída, entre las antenas y las mandíbulas. A la derecha, los abanicos se hallan extendidos, con sus radios abiertos para filtrar las partículas alimenticias del agua. Las partículas se adhieren a un recubrimiento de mucus situado sobre los radios, y la larva se alimenta acercando los radios a su boca y separando luego las partículas con sus mandíbulas. Ambas fotografías fueron realizadas por Douglas A. M. Craig, de la Universidad de Alberta.



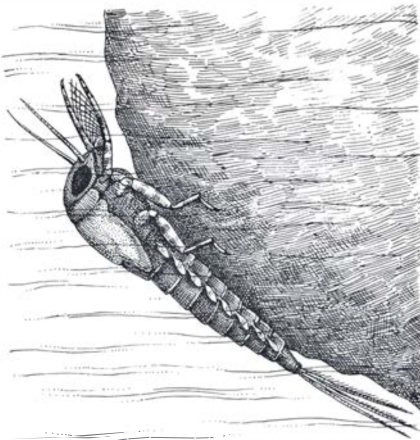


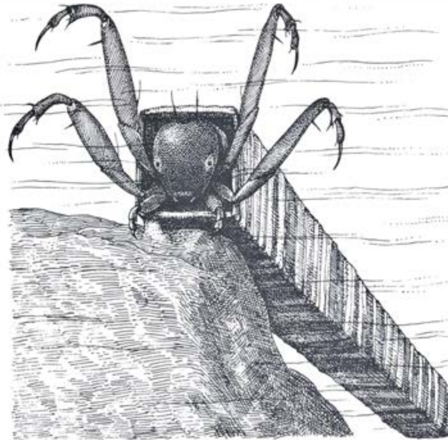
DISPOSICION EN ABANICO de las sedas erizadas sobre la pata anterior de una ninfa del género *Isonychia*, un efemeróptero, según se aprecia en la micrografía electrónica de barrido de la derecha. A la izquierda, a mayor aumento, se advierten los minúsculos pelos situados sobre cada seda; la mitad de ellos son cortos y ganchudos; la otra mitad son más largos y curvados. Los pelos ganchudos de una seda se traban con los pelos largos de la siguiente para formar un filtro que puede atrapar, cuando el animal se sitúa contracorriente, partículas alimenticias de un diámetro inferior a un micrometro.

cionados al comienzo de este artículo: los que hilan redes. Empecemos por los tricópteros de la familia Hidropsíquidos. Las larvas de determinados miembros de este grupo construyen en primer lugar un refugio de restos orgánicos e inorgánicos que unen entre sí con su propia seda. El extremo abierto de la casita así fabricada está dirigido corriente arriba. A continuación, la larva construye un armazón en forma de aro oval que sostendrá una red para capturar alimento en el extremo abierto del refugio. Después se hila la red, empezando cerca de la base del armazón, mientras la larva hace oscilar su cabeza en una serie de movimientos que siguen el trazado de un ocho.

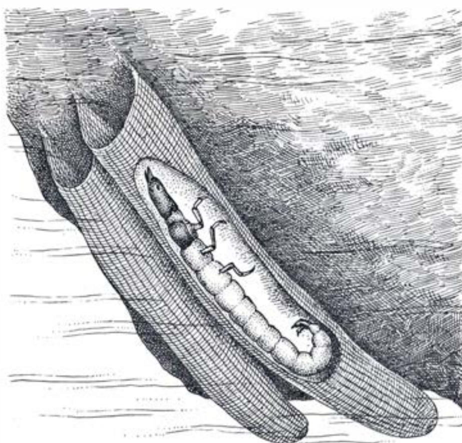
La primera hebra de pegajosa seda se tiende diagonalmente desde un lado del armazón hasta la base. La segunda hebra se tiende de igual manera, desde el otro lado. La alternancia continúa, corriendo paralelas entre sí las hebras de cada lado. El resultado final es una red con una malla rectangular y una costura central que divide una mitad de la otra. El entomólogo alemán Werner Sattler, que había estudiado el género de tricópteros *Hydropsyche,* observó que la larva precisaba de siete a ocho minutos para hilar su red. Si la red se desgarraba, la larva la remendaba al azar; si la red resultaba fuertemente dañada, la larva tejía otra nueva.

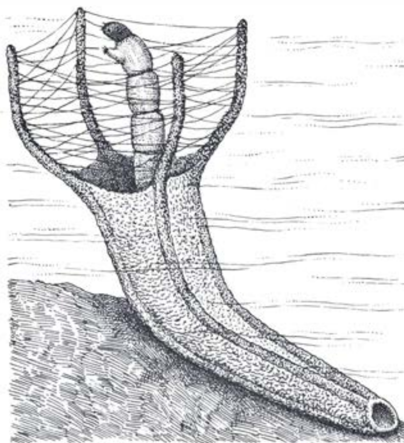
Las larvas de tricópteros mudan varias veces antes de alcanzar el estadio adulto. Las redes de captura construidas después de cada muda son progresivamente mayores y de malla más gruesa, y, más robustas, sus hebras. Aunque las redes con mallas más gruesas son menos eficientes en la captura de partículas pequeñas, no sólo son mayores, sino que con frecuencia se sitúan en lugares donde la corriente es más rápida. De ahí que filtren un volumen de agua superior que una red de malla fina colocada en una corriente más lenta.





PATAS MODIFICADAS que la ninfa del género de efemerópteros *Isonychia,* a la izquierda, y la larva del género de tricópteros *Brachycentrus,* a la derecha, utilizan para capturar partículas alimenticias. La ninfa de efímera se sitúa contra la corriente y utiliza los pinceles de sus piezas bucales para peinar las sedas de sus patas anteriores y limpiarlas de las partículas alimenticias atrapadas. La larva de tricóptero construye un refugio móvil que ocupa después de fijarlo, orientado corriente arriba. Las series de sedas cortas filtradoras se hallan dispuestas en las patas posteriores y medias, que limpia con sus patas anteriores.

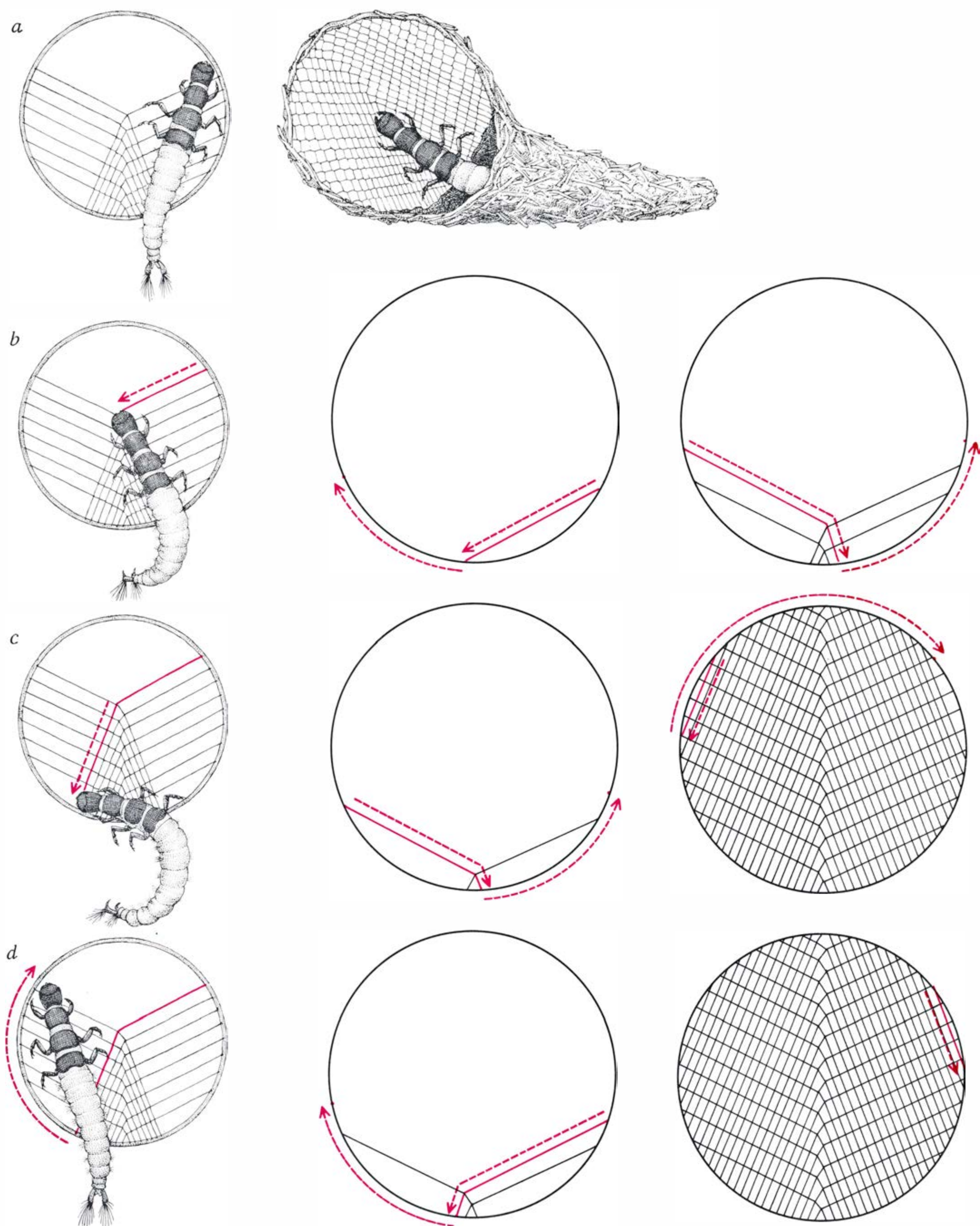




REDES SIMPLES, construidas por la larva de un tricóptero de la familia Filopotámidos, a la izquierda, y la larva de un díptero del género *Rheotanytarsus,* a la derecha. La larva de tricóptero construye una red larga y sacciforme, con una malla muy fina, fijada de manera que la abertura mayor quede encarada corriente arriba. La larva ocupa la red y periódicamente utiliza el pincel de su labio superior para extraer las partículas alimenticias atrapadas sobre la superficie interna de la red. La larva de díptero construye una guarida tubular a base de partículas de sedimento unidas mediante su saliva siriciforme y añade unos brazos que surgen del extremo dirigido hacia la corriente. Después, tiende unos filamentos pegajosos entre los brazos; por lo general come los filamentos y las partículas de alimento que se adhieren a ellos.

Theodore J. Georgian, Jr., de la Universidad de Georgia, y uno de nosotros (Wallace) han desarrollado un modelo de captura de partículas por parte de tricópteros constructores de redes en un arroyo de los Apalaches meridionales. El modelo sugiere que las larvas de todas las especies de tricópteros, independientemente de su estadio de madurez y del tamaño de malla de sus redes, se las arreglan para filtrar más alimento del que necesitan. Sin embargo, la mayor parte de lo que capturan son detritos, es decir, material alimenticio de baja calidad. El análisis del contenido intestinal de varias especies de tricópteros indica que las que tienen redes de malla gruesa se alimentan sobre todo de partículas de tejido animal, relativamente grandes y relativamente escasas, mientras que las que poseen redes de malla fina se alimentan principalmente de detritos de tamaño de partícula menor, que son más abundantes. De ahí que la ventaja de que un gran volumen de agua pase a través de una red de malla gruesa, en vez de un pequeño volumen a través de una red de malla fina, sea que la red gruesa puede ser más selectiva para los alimentos animales. Las diferencias observadas en el tamaño de malla de las redes de larvas de tricópteros de distintas especies y en diferentes estadios de crecimiento parecen estar más relacionadas con la selección de distintos tipos de alimento que con la selección de partículas de un determinado tamaño.

Después de haber capturado las partículas alimenticias, grandes o pequeñas, en redes de malla gruesa o fina, ¿cómo las ingieren las larvas de tricópteros? Al igual que los filtradores que no tejen redes, los que las construyen han desarrollado distintas adaptaciones estructurales y de comportamiento. Por ejemplo, una de las mallas más gruesas tejida por un miembro de la familia Hidropsíquidos es la del género *Arctopsyche,* que vive de preferencia en aguas de corrientes rápidas. Las larvas de este género suelen capturar presas vivas en



RED COMPLEJA CONSTRUIDA POR LAS LARVAS de los tricópteros del género *Hydropsyche*. Se muestra arriba con su armazón en forma de aro situado en la entrada de la guarida subacuática de la larva. La secuencia de la izquierda muestra el método de construcción, descrito por Werner Sattler. La larva sujeta el armazón y la redecilla y fija un hilo de seda al borde derecho del armazón. Primero extiende el hilo (*color*) al centro de la red, y luego hacia abajo hasta el borde del armazón, en el otro lado. Una oscilación hacia arriba y a la izquierda y una repetición especular del primer movimiento completan el movimiento en forma de ocho de la larva y añaden otra hebra a la red. Los diagramas esquemáticos del centro y de la derecha señalan la fijación de los primeros cuatro filamentos y de los dos últimos. La larva de *Hydropsyche* puede construir su red en siete u ocho minutos. (Ilustración de Tom Prentiss.)

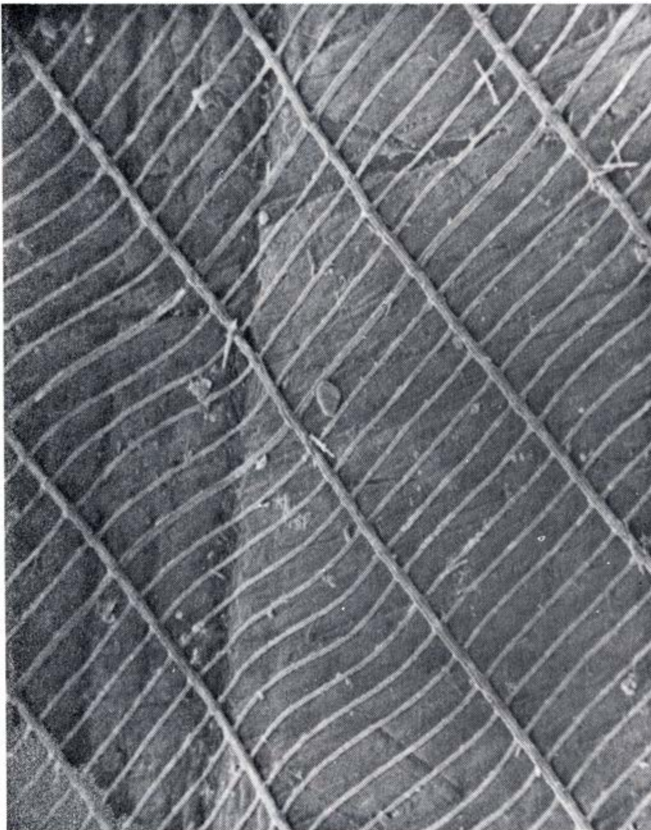
sus redes, y sus patas anteriores espinosas sirven a tal fin. Las larvas del género *Macronema* viven en aguas más tranquilas e hilan redes con una malla muy fina a lo largo de todos sus sucesivos estadios de crecimiento. Sus patas anteriores y sus piezas bucales están equipadas con densas series de sedas; con estos "pinceles" las larvas capturan e ingieren las partículas de alimento que se acumulan en la red. Las larvas de otros dos géneros de tricópteros, *Phylocentropus* y *Protodipseudopsis,* recolectan asimismo las partículas alimenticias de sus redes mediante las patas anteriores, que portan pinceles.

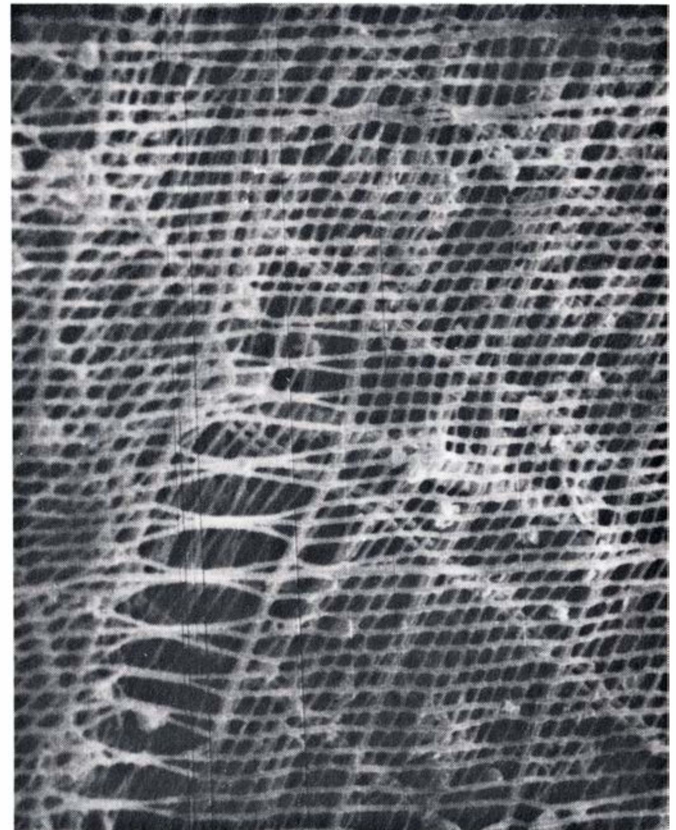
Algunas larvas de tricópteros que hilan redes capturan el alimento de otras maneras. Las de la familia Filopotámidos construyen un tubo sacciforme que sirve a la vez de cobijo y de red captadora. Estos sacos destacan por dos motivos. En primer lugar, por su magnitud: llegan a medir hasta cinco centímetros de longitud y tres de diámetro. En segundo lugar, por la finura extraordinaria de su malla. Por ejemplo, cada una de los 10 millones de aberturas rectangulares de la malla del saco que construyen las larvas del género *Dolophilodes* en su estadio larvario final miden 0,5 micrometros por 5,5 micrometros. Las aberturas de la malla del género *Wormaldia,* formada por capas superpuestas de malla rectangular, miden sólo 0,4 por 0,4 micrometros. Los cobijos de malla fina tienen una abertura grande en el extremo dirigido contra la corriente y una pequeña abertura en el extremo situado corriente abajo. En el interior del cobijo, la larva de filopotámido barre periódicamente las partículas de detritos finos de la malla y las dirige hacia su boca, utilizando las cerdas de su labio superior.

La larva del pequeño jején *Rheotanytarsus,* un filtrador del orden Dípteros, construye una guarida en forma de tubo sobre la superficie de piedras o de restos vegetales sumergidos. La estructura está formada por partículas de sedimento unidas mediante la saliva siriciforme de la larva; la abertura que mira corriente arriba es grande y la opuesta, reducida. A esta estructura la larva añade de dos a cinco brazos delgados que salen hacia arriba desde el extremo grande, y entre los brazos tiende varios filamentos para hacerse en esta trampa con los detritos que pasen. De vez en cuando, la larva sale en parte de su cobijo, se come los filamentos junto con cualesquiera partículas alimenticias que se hayan adherido a ellos y fija una nueva dotación de hebras.

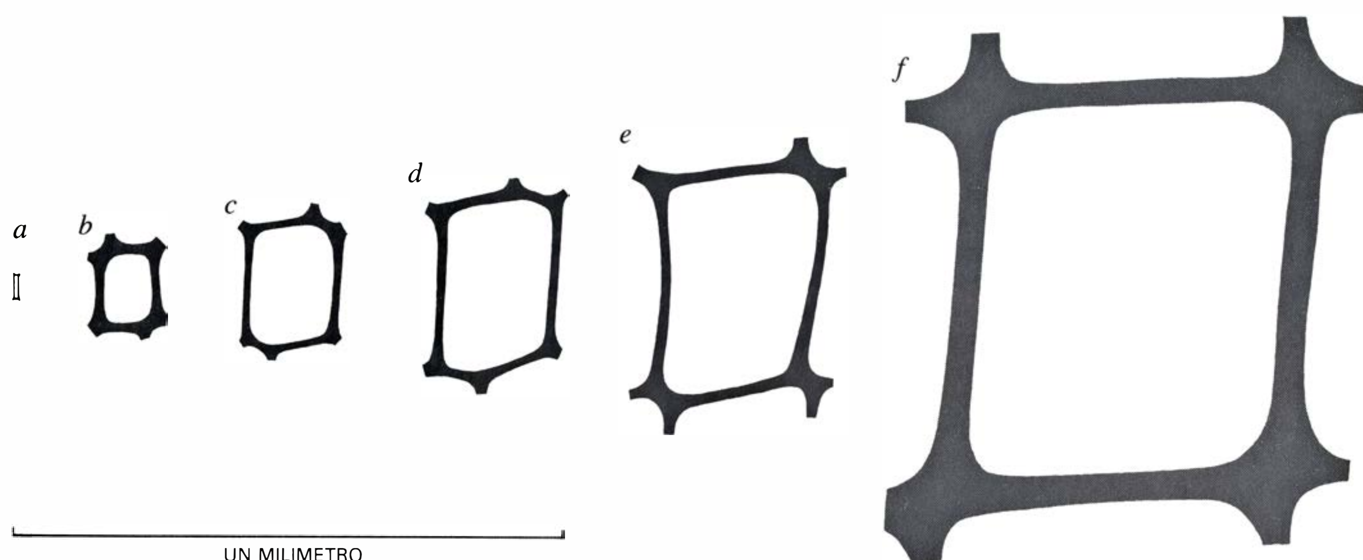
Varios tricópteros, efemerópteros y dípteros filtradores pasan sus estadios inmaturos en aguas estancadas o de corriente lenta. De ellos, los más conocidos son los dípteros de la familia Culícidos, es decir, los mosquitos. Las larvas de mosquito, de las que se conoce el activo movimiento de agitación de su cuerpo, se alimentan de materia orgánica suspendida en el agua mediante un par de pinceles bucales modificados. Los pinceles son similares a los abanicos cefálicos de las larvas de simulíidos en su estructura, musculatura y función. No es preciso que las aguas que ocupan las larvas de mosquito estén limpias ni que sean de gran extensión. Se encuentran larvas de mosquito en huecos de troncos de árboles, en charcas de nieve e incluso en las gotitas de agua que se acumulan en la base de una hoja; también se encuentran en agua salobre y en pozos de letrinas. Pueden medrar en hábitats estacionarios, como los indicados, porque el movimiento de sus pinceles bucales genera pequeñas corrientes que ponen a su alcance partículas alimenticias.

Algunos grupos de la familia de los mosquitos no son filtradores, sino que





MALLAS DE DOS TAMAÑOS, ambas tejidas por larvas de tricópteros. A la izquierda se advierte la malla oblonga del género de hidropsíquidos *Macronema;* cada poro mide unos cinco micrometros de anchura y 40 de longitud. A la derecha se muestra la malla en doble capa del género de filopotámidos *Wormaldia*. La superposición de dos mallas rectangulares produce el poro de red más pequeño que se conoce en los tricópteros: 0,4 por 0,4 micrometros.



TAMAÑOS DE MALLA, que difieren mucho entre los distintos géneros de tricópteros hidropsíquidos. Los diversos poros que se muestran aquí esquemáticamente van desde el oblongo de *Macronema* (*a*) hasta el rectangular de *Arctopsyche* (*f*). Los seis tipos de mallas son representativos de las redes que las larvas de tricópteros de la cuenca del río Savannah construyen en su estadio larvario final. Las larvas de *Arctopsyche* prefieren corrientes rápidas.

han desarrollado piezas bucales modificadas para el ramoneo. Mientras que las larvas filtradoras tienden a recolectar las partículas justo debajo de la superficie del agua, los ramoneadores se alimentan generalmente sobre el fondo. Sus piezas bucales modificadas les permiten raspar partículas alimenticias de los restos orgánicos del fondo. La mayoría de larvas que viven cerca de la superficie capturan e ingieren partículas de menos de 50 micrometros de diámetro.

Distintos estudios muestran que la supervivencia de las larvas de mosquito no sólo se halla regulada por factores ambientales principales, como el fotoperíodo y la temperatura, salinidad y concentración de oxígeno del agua en la que habitan, sino también por factores químicos sutiles. Rex H. Dadd, de la Universidad de California en Berkeley, ha demostrado que algunos de estos factores químicos reguladores aceleran el crecimiento de las larvas al aumentar la tasa de ingestión de alimento y el tiempo que las larvas pasan alimentándose. El efecto último de esta estimulación química sería acelerar la formación de densas poblaciones larvarias, si no fuera porque, cuando las larvas más maduras crecen en condiciones de hacinamiento, producen sustancias que son muy tóxicas para las menos maduras.

Al igual que las larvas de mosquito, las larvas hiladoras de seda de la familia Quironómidos prefieren los lagos y otras aguas estancadas o de corrientes lentas. Pueden excavar una madriguera en el sedimento del fondo o fijar su cobijo sobre la superficie de un tronco sumergido o el tallo o la hoja de una planta acuática. La capacidad de hilar de las larvas de jejenes les ha permitido adaptarse a un amplio espectro de hábitats. Constituyen uno de los consumidores primarios más importantes en las cadenas tróficas acuáticas y en ocasiones alcanzan densidades de población de más de 50.000 individuos por metro cuadrado de fondo.

Las larvas de quironómidos pueden excavar simplemente una galería en el blando sedimento lacustre o bien pueden trabajar con su seda para construir un cobijo a partir de la materia particulada de que disponen. La larva teje una tenue red cónica a través de la boca del cobijo, tarea que le ocupa unos 30 segundos y, luego, mientras ondula su cuerpo, bombea agua a través de la red y del refugio a la vez. Si una acumulación de detritos tapona la red, la larva invierte sus ondulaciones, generando una fuerte contracorriente que termina por desatascar la obstrucción.

Al objeto de alimentarse, la larva de quironómido se fija firmemente al revestimiento de seda de su galería mediante garras ganchudas y devora a la vez la red y las partículas alimenticias adheridas a ella. En el intervalo comprendido entre la ingestión de la vieja red y la construcción de la nueva, la larva defeca. El tiempo total empleado en el ciclo, incluyendo los 30 segundos del intervalo de hilado de la red, es de tres a cuatro minutos. Salvo en que las larvas de quironómidos crean su propio flujo de agua y comen su red en lugar de limpiarla extrayendo de la misma las partículas alimenticias que se adhieren a ella, desempeñan en la cadena trófica de las aguas lentas y de deposición un papel equivalente al de las larvas tejedoras de redes de los tricópteros en las aguas rápidas y de erosión.

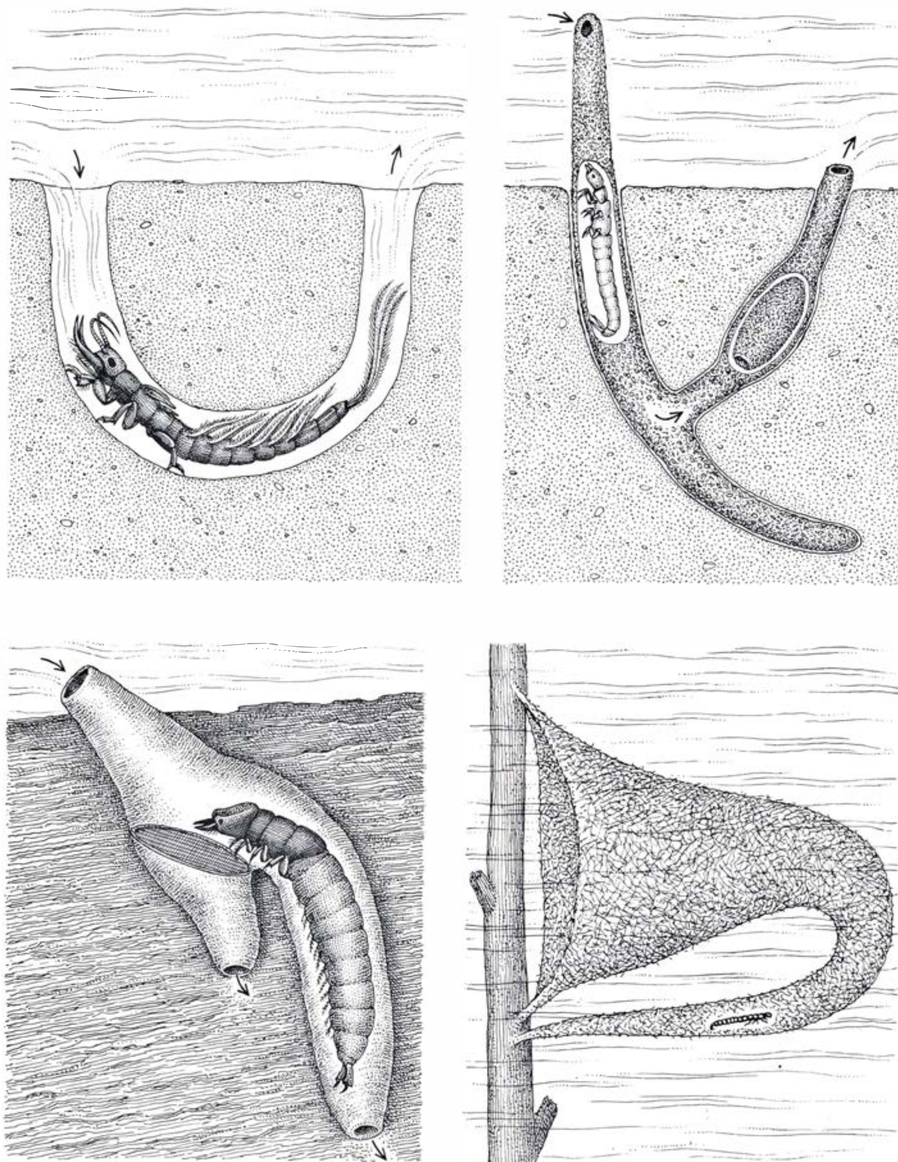
Las ninfas de algunos géneros de efemerópteros viven en el fango y limo de los fondos cercanos a la orilla en lagos y ríos de aguas lentas; también pueden vivir en madera sumergida, como troncos de árboles y pilotes de desembarcaderos. Dos organismos bentónicos (que viven en el fondo) comunes son las ninfas de los géneros *Hexagenia* y *Ephemera;* ambos géneros se sirven de patas modificadas para excavar y construir una madriguera en forma de *U* en el sedimento del fondo. Una vez instalada la ninfa en su galería, comienza a ondular sus branquias respiratorias. La corriente generada por el movimiento aporta a la vez agua rica en oxígeno y partículas alimenticias a la madriguera. Varios estudios sugieren que las ninfas pueden alimentarse en sus galerías filtrando partículas del agua que las atraviesa. Sin embargo, también se sabe que salen de su refugio y ramonean sobre el fondo. En muchas zonas del medio oeste de los Estados Unidos, miríadas de estas efímeras cubrían antaño el suelo durante su breve emergencia estival. En la actualidad, su número se ha reducido mucho debido a la contaminación de las aguas donde habitan.

Las ninfas de dos géneros de efímeras tropicales, *Povilla* y *Astenopus,* se cuentan entre las que atacan madera sumergida. Excavan una madriguera en

forma de *U* en la madera mediante sus fuertes piezas bucales y se refugian en ella después de revestirla de un material sedoso. Ondean luego sus largas branquias abdominales para generar un flujo de agua que transporta oxígeno y partículas alimenticias a través de la madriguera. Filtran el alimento del agua circulante mediante densas series de sedas de sus patas anteriores, cabeza y piezas bucales.

Entre las larvas de tricópteros que viven de modo similar en lagos y arroyos lentos se cuenta la larva de un género, *Neureclipsis,* que teje una red con una forma característica de cornucopia. La entomólogo alemana Caroline Brickenstein encontró, estudiando estas larvas, que tardaban tres días en construir su red. Las redes mayores pueden tener más de 20 centímetros de longitud y una abertura de 13 centímetros de diámetro. La seda pierde buena parte de su elasticidad y solidez a los pocos días, y la larva invierte un tiempo adicional entre los períodos de alimentación sustituyendo las hebras gastadas de la pared interna.

La larva de *Neureclipsis* se alimenta, sobre todo, de pequeños invertebrados acuáticos, de modo que, a menudo, sus redes se encuentran en gran abundancia en los ríos que desaguan lagos en los que estas minúsculas presas existen en gran número. Aunque el hábitat preferido de las larvas son las aguas tranquilas, a veces construyen sus redes en aguas que fluyen a velocidades de hasta 30 centímetros por segundo e incluso más. Las redes que ocasionalmente pueden verse en estos ríos de aguas rápidas son claramente mucho más pequeñas que las que se hallan en aguas lentas, lo que implica que las velocidades superiores imponen limitaciones al tamaño de la red.



FILTRADORES DE AGUAS LENTAS, de los que aquí se ilustran dos, una ninfa del género de efemerópteros *Hexagenia*, arriba a la izquierda, y una larva del género de tricópteros *Phylocentropus*, arriba a la derecha; ocupan tubos o madrigueras y bombean corrientes de agua a través de sus refugios haciendo ondular el cuerpo. La ninfa de efímera establece el flujo de agua haciendo oscilar sus branquias abdominales dorsales y recoge las partículas alimenticias con los pinceles de sedas que se hallan sobre sus patas anteriores y sus piezas bucales. La larva de tricóptero construye su tubo ramificado con seda y granos de arena y teje una red irregular en la rama más corta; los movimientos del cuerpo atraen el agua hacia la rama larga y la empujan por la corta y hacia el exterior; la larva barre periódicamente la red y limpia la pared interna del tubo con los pinceles de sus patas anteriores y piezas bucales. Otros dos tricópteros utilizan corrientes suaves para el aporte de partículas alimenticias a sus redes. La larva del género *Macronema*, abajo a la izquierda, construye un refugio con su entrada dirigida corriente arriba y limpia su red con los pinceles de sus patas anteriores y sus piezas bucales. La larva del género *Neureclipsis,* abajo a la derecha, construye una gran red en forma de cornucopia que puede tener 20 centímetros de longitud. La larva se alimenta generalmente de presas vivas que se acumulan en el extremo estrecho de la red.

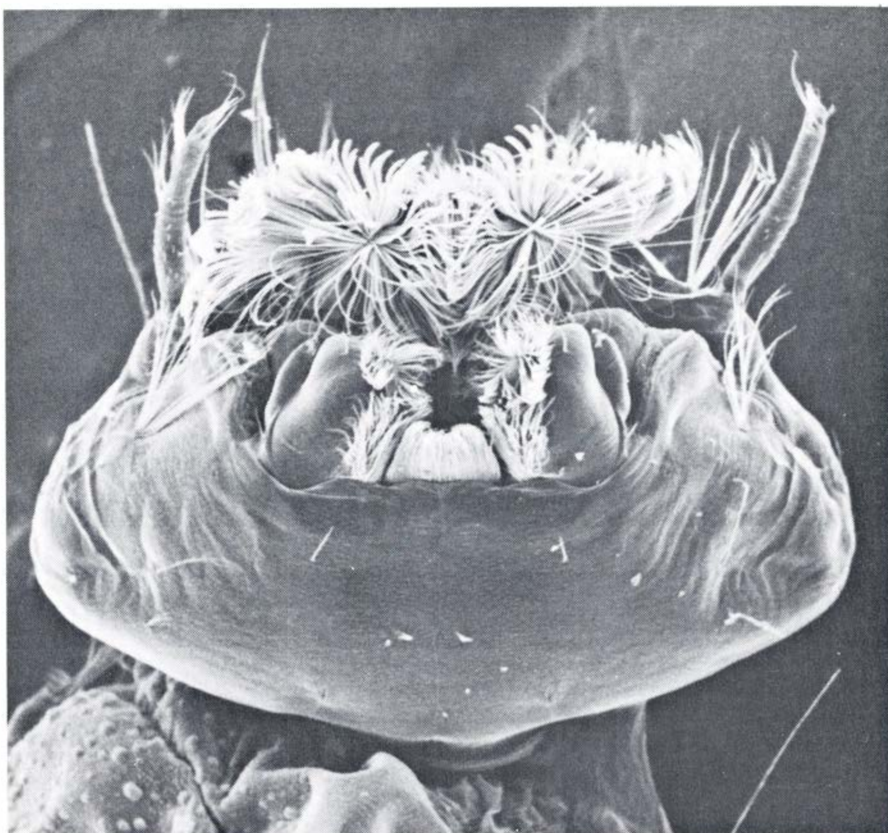
La larva del género de tricópteros *Phylocentropus* es una de las pocas frigáneas hiladoras de red que viven en zonas fluviales donde la deposición predomina sobre la erosión. Construye un largo tubo en forma de *Y* que se halla enterrado a varios centímetros de profundidad en el fondo del río. Un brazo de la *Y* es alargado y se extiende varios centímetros hacia arriba, en el agua; el otro brazo, que tiene una protuberancia, es corto y apenas sobresale del nivel del fondo. La larva ocupa normalmente el brazo más largo. Al ondular su cuerpo, provoca que el agua fluya hacia el brazo más largo y salga por el más corto, en el que el abultamiento contiene una red captadora. Entre los intervalos de ondulación, la larva penetra en el brazo corto de la madriguera y se alimenta del finísimo detrito que se adhiere a la red y a las paredes internas del tubo. Los análisis del contenido intestinal indican que la mayoría de las partículas que ingieren las larvas de *Phylocentropus* tienen un diámetro de menos de 10 micrometros.

Los hábitats disponibles en un determinado ecosistema acuático tienen un número limitado. ¿Cómo consiguen compartirlos los insectos filtradores? La respuesta es evidente: los distintos géneros han desarrollado muchos mecanismos adaptativos diferentes, tanto de comportamiento como estructurales. Un mecanismo ulterior, que supone la selección del hábitat, es común entre determinados filtradores, en particular las larvas de simulíidos y las larvas hiladoras de los tricópteros de la familia Hidropsíquidos.

El agua de un lago o de un embalse

contiene productos de descomposición (procedentes de los sedimentos bentónicos) y grandes poblaciones de vegetales y animales microscópicos. En consecuencia, estas larvas se congregan en gran número cerca de los desaguaderos del lago y de los aliviaderos de la presa. Con frecuencia, la época del año en la que el agua rica en nutrientes vierte río abajo es primavera, de modo que los filtradores inmaturos situados cerca de un desaguadero en esa época gozan de una ventaja selectiva sobre los insectos que maduran en otras estaciones del año o en hábitats situados corriente abajo. Se ha encontrado, como cabía esperar, que algunas especies de insectos que ocupan estos hábitats caracterizados por aguas ricas en nutrientes crecen más rápidamente y alcanzan antes el estado adulto que los representantes de la misma especie que viven en otros lugares. El resultado es un ciclo biológico abreviado, mejor adaptado a la explotación de una abundancia estacional de alimento.

La tendencia de las larvas de simulíidos a congregarse en estos hábitats de vertidos plantea un grave problema sanitario en Africa. La picadura de la hembra adulta de *Simulium* transmite las filarias que en el hombre causan la oncocercosis, la llamada "ceguera fluvial". La construcción de embalses y otras represas en las naciones en vías de desarrollo de Africa ha conducido a un aumento en el número de zonas de cría de los simulíidos y a una expansión de la enfermedad.



PINCELES BUCALES de una larva de mosquito, vistos desde abajo en esta micrografía electrónica de barrido realizada por Craig. El mosquito ilustrado (*Culiseta inornata*) es un filtrador de hábitats de aguas estancadas, y mueve los pinceles rítmicamente para atraer una corriente con alimento hacia su boca.

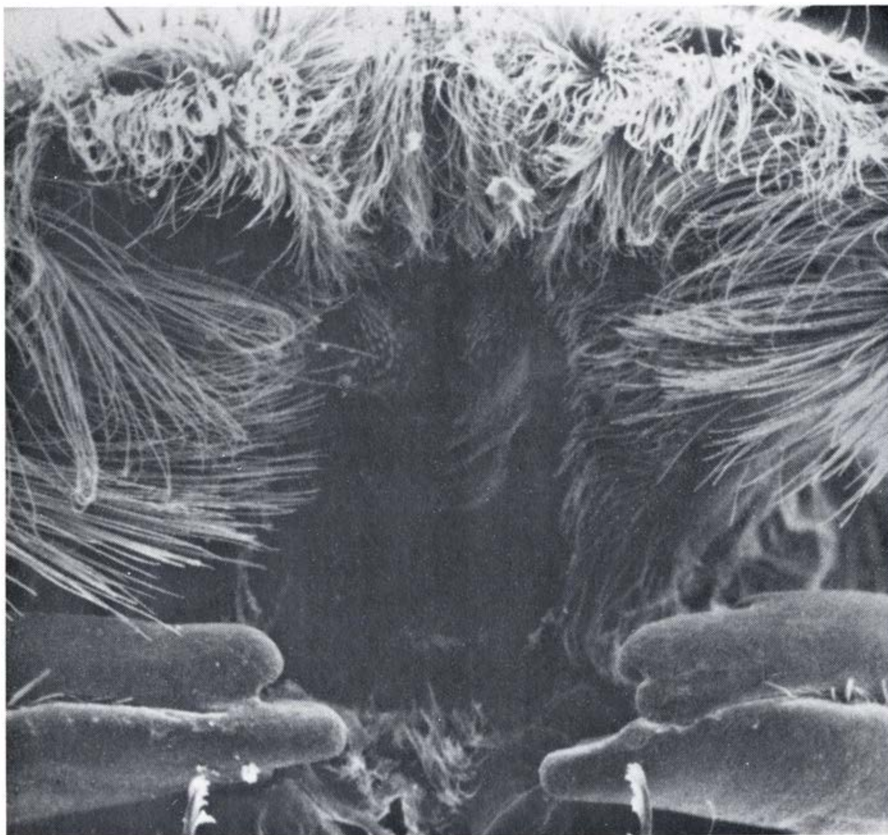
Otros factores distintos de la abundancia de alimento afectan a la duración del ciclo biológico de un filtrador. La temperatura del agua o, más exactamente, la acumulación de calor a lo largo de un período determinado de tiempo, es uno de tales factores. Trabajando con Ross, en Michigan, uno de nosotros (Merritt) encontró que las larvas de simulíidos tardaban más en desarrollarse cuando hacían eclosión a temperaturas invernales del agua cercanas a la de congelación que cuando salían del huevo y se desarrollaban a las temperaturas en aumento del agua primaveral. Las larvas que se desarrollan en invierno pasan asimismo a través de más mudas y son mayores cuando inician el estadio pupal que las larvas que se desarrollan más rápidamente en primavera. En Georgia, uno de nosotros (Wallace) ha observado una respuesta algo similar a la temperatura del agua entre las larvas de tricópteros en maduración, aunque las temperaturas más bajas no producen mudas adicionales. Otras investigaciones han demostrado que los insectos filtradores que viven a altitudes mayores o en latitudes más frías producen normalmente una generación por año, mientras que las que son propias de altitudes menores o de latitudes más cálidas pueden producir dos o incluso tres generaciones en el mismo período de tiempo.

Otro factor adaptativo en la coexistencia de los filtradores es un escalonamiento del ciclo biológico, como ocurre cuando los insectos ocupan el mismo hábitat, pero no pasan a través de las mismas fases de desarrollo al mismo tiempo. Estas variaciones temporales en el ciclo biológico pueden ser útiles para varios fines adaptativos. Por ejemplo, en un momento determinado una población puede estar consumiendo alimentos que son distintos de los de otra. O bien el período de máximo crecimiento en una población puede diferir del de la otra, de modo que los momentos de máxima demanda de alimento se hallan escalonados.

Existen asimismo cambios en la dieta con el desarrollo larvario. Varias investigaciones han demostrado que las larvas tempranas de algunos tricópteros hidropsíquidos se alimentan principalmente de detritos finos y de diatomeas y otras algas. Sin embargo, al pasar por mudas sucesivas, las larvas empiezan a consumir cantidades crecientes de tejidos animales. El alimento que el filtrador ingiere de preferencia durante su período de crecimiento máximo es el que puede asimilar de manera más eficiente.

Aldo S. Leopold, un pionero en los estudios de la fauna silvestre, señaló una vez que los procesos de la naturaleza hacen que todos los materiales, los orgánicos incluidos, se muevan, de modo predominante, cuesta abajo. Leopold sugirió que la continuidad y la estabilidad de las comunidades de las tierras altas quizá dependiera de los tipos de organismos que almacenan nutrientes y materia orgánica y participan en otros procesos que retardan la tendencia a la pérdida cuesta abajo. Es evidente que los insectos acuáticos filtradores se hallan entre estos tipos de organismos, pero es difícil medir con precisión su contribución al retardo indicado.

La dificultad surge del flujo de agua unidireccional, que es el principal portador de materiales corriente abajo que



PINCELES BUCALES DE UN TRICOPTERO, usados para limpiar las finas partículas de su red captadora y llevarlas a su boca. En esta micrografía electrónica de barrido se ve la larva de un tricóptero hidropsíquido del género *Macronema*. Los objetos de la parte baja de la micrografía son las mandíbulas.

Leopold imaginaba. Con un sistema abierto, como lo es un río que fluye, el valor relativo de las entradas y de las salidas no resulta fácil de calcular. No obstante, son posibles algunas evaluaciones de sentido común. Por ejemplo, en comparación con el transporte total corriente abajo existe un pequeño movimiento de materiales corriente arriba. Pueden citarse casos de movimiento neto corriente arriba realizado por peces y por animales bentónicos, e incluso por insectos acuáticos después de haber alcanzado la madurez y emprendido el vuelo, pero todos estos movimientos son irrelevantes comparados con el que se hace corriente abajo.

¿Qué es lo que consiguen exactamente los filtradores al retardar el proceso de caída cuesta abajo? Una de sus principales aportaciones puede ser el retener parte del alimento que ingieren y alterar el resto y hacerlo circular. En este sentido, estudios de seis especies de larvas de tricópteros tejedoras de redes de un río de los Apalaches meridionales indican que estos filtradores añaden de hecho más detritos a la corriente de los que extraen de ella. Sólo del 2 al 20 por ciento de la ingesta de detritos que realiza el filtrador se asimila. Al mismo tiempo, la larva se halla ingiriendo, pero en absoluto asimilando por completo, tejido animal de elevada calidad y material vegetal de calidad algo menor. Las heces de los filtradores, aunque pueden contener entre el 80 y el 98 por ciento de la propia ingesta de la larva de detritos de baja calidad, encierran asimismo algún material animal y vegetal no asimilado. A través de la asimilación, las larvas reducen el valor alimentario neto de lo que ingieren, pero sus heces, junto con aquellos microorganismos colonizadores que las heces pueden adquirir en su recorrido, se hallan en disposición de ser reingeridas por otros filtradores situados en el curso inferior del río.

Este mecanismo puede considerarse punto de partida de un proceso de reciclado que aumenta la eficiencia de un ecosistema fluvial en términos de la utilización de sus entradas orgánicas. Si no fuera por los filtradores, gran parte de la materia orgánica que se transporta corriente abajo sería utilizada sólo por los componentes microbianos del ecosistema. Sin embargo, en un río donde hay sucesivas poblaciones activas de filtradores, la materia orgánica puede rendir de manera repetida fracciones de su energía almacenada en su largo viaje hasta el mar.

Jackson R. Webster, del Instituto Politécnico y la Universidad estatal de Virginia, ha bautizado este proceso de reciclado con el término "espiralización". Lo ha hecho así para resaltar los aspectos longitudinales y unidireccionales del ciclado dentro de un ambiente fluvial. Por ejemplo, la materia particulada de alta calidad, como los tejidos animales y las algas, puede utilizarse rápidamente; puede caracterizarse, pues, por poseer una corta distancia de espiralización. Los detritos de baja calidad pueden definirse por tener una distancia de espiralización mayor. Cuanto más corta sea la distancia de espiralización, mayor será la proporción de materia orgánica que se convierte en dióxido de carbono por el metabolismo de los filtradores, y de ese modo se sustrae de la cantidad total de materia orgánica que viaja corriente abajo. Y cuanto más diversos sean los mecanismos de captura de los distintos filtradores, mayor será la eficiencia con la que se elimina la materia orgánica.

Sin embargo, éste es sólo uno de los factores que se relacionan con la eficiencia de la utilización del alimento. Por ejemplo, la materia particulada que es más abundante en los ríos tiene un diámetro de partícula de 25 micrometros o menos. Los insectos filtradores que capturan partículas de este rango de tamaño son simulíidos, quironómidos y algunos tricópteros. Son los mismos filtradores que seleccionan el alimento casi enteramente por el tamaño de partícula. Por tanto, en la eficiencia global del ecosistema estos insectos pueden conseguir más en la retención de materia orgánica que los filtradores que atrapan selectivamente partículas mayores.

Los pocos estudios que se han hecho hasta ahora indican que, sobre distancias cortas corriente abajo, los insectos filtradores emplean únicamente una pequeña proporción de la materia orgánica circulante. Sin embargo, convierten esta proporción en materia orgánica de una forma más compleja y con un mayor valor alimentario. Esta materia orgánica está constituida por su propio cuerpo, que es un alimento potencial para los depredadores, como los insectos carnívoros y los peces situados más arriba en la cadena trófica. De ahí resulta claro que los filtradores, al haber evolucionado para ocupar distintos hábitats y para emplear muchos tipos de mecanismos de captura, actúan retardando el movimiento dominante de la materia orgánica, corriente abajo, reteniéndola y alterándola a la vez.

Naves de guerra a remo en la antigüedad

La historia de estas naves, extraída de diversas fuentes, permite deducir cómo evolucionaron y se adaptaron a las nuevas exigencias derivadas del progreso de la táctica militar

Vernard Foley y Werner Soedel

En la época de apogeo de Grecia y Roma, las naves de guerra propulsadas a remo desempeñaron un destacado papel en el mantenimiento del tráfico comercial a larga distancia y de los vínculos de unión del imperio. La vela, que había aparecido hacia el año 3500 a. de C., la utilizaban fundamentalmente las naves mercantes, por cuanto éstas necesitaban operar con unos costos de explotación relativamente bajos y, en consecuencia, llevaban unas tripulaciones reducidas al mínimo; por cuyo motivo empleaban los remos solamente en las maniobras de entrada y salida del puerto, o para llegar a algún punto próximo cuando el viento cesaba de soplar. Las naves de guerra navegaban impulsadas por el viento todo cuanto podían, pero al acercarse el momento de la batalla desguarnían los palos y las velas dejándolos en cualquier playa próxima. La fuerza producida normalmente a base de remos era bastante menor que la del viento. Sin embargo, en períodos de tiempo relativamente cortos, bastaba para imprimir a las naves de guerra una velocidad muy adecuada, y al mismo tiempo proporcionaba una notable capacidad de maniobra. El metabolismo de los remeros actuaba de depósito de energía, que liberaban en un momento dado de forma muy rápida, cosa que no ocurría con las velas.

Pese a las limitaciones de la propulsión a remo en lo que concierne a largas navegaciones, lo cierto es que las naves de remo, es decir, las galeras de la antigüedad, llegaron a ser un verdadero alarde de ingeniería. La nave de guerra griega por excelencia, en la época de su esplendor, era el trirreme, nombre con el que se la conoce normalmente en castellano. La palabra en sí procede de la voz latina *triremis* y ésta, a su vez, de la griega *trieres,* que de un modo muy aproximado podríamos traducir por "equipada a tríos". Hasta hace pocos años, el significado exacto de la palabra fue motivo de polémica entre los eruditos, pero actualmente la cuestión ha quedado totalmente resuelta gracias a los trabajos del británico J. S. Morrison, quien de forma magistral la ha estudiado y descrito basándose en las citas literarias, las inscripciones lapidarias, las representaciones gráficas y las medidas obtenidas en los lugares donde se varaban.

El trirreme fue el resultado de la evolución de unas embarcaciones muy sencillas y de origen fundamentalmente griego y fenicio. Se trataba de unas embarcaciones sin cubierta, propulsadas a remo y con una fila, u orden de remeros, por banda. A partir de la decoración de cerámica y de algunos pasajes de Homero se deduce que estaban concebidas para desarrollar una velocidad elevada, aprovechando lo mejor posible la energía muscular de los remeros. De todos modos no alcanzaron toda su eficacia hasta la incorporación de un nuevo elemento: el espolón, que apareció hacia el 800 a. de C. y produjo una verdadera revolución en el sector de la construcción naval. Hasta entonces, los combates navales habían consistido siempre en una lucha cuerpo a cuerpo, que se iniciaba una vez producido el abordaje entre las embarcaciones. A partir de la introducción del espolón, en cambio, fue posible hundir o destruir la nave enemiga, en cuyo caso el herir o dejar fuera de combate a la tripulación pasó a ser algo totalmente secundario.

El empleo del espolón exigía una mayor velocidad y maniobrabilidad de las embarcaciones. Por esa razón, las de la época anterior a Homero evolucionaron con rapidez hasta convertirse en unas naves de mucha eslora, poco puntal y formas muy finas, propulsadas por una fila de veinticinco remeros a cada banda, y que se designaban con el nombre de *pentecóntoras,* o galeras de cincuenta remos. Las proporciones del casco de estas naves, que se conocen con bastante exactitud, se mantuvieron posteriormente en las galeras de estructura más complicada.

La resistencia a la marcha de un buque en el agua es la suma de cuatro componentes principales. Uno de ellos es la resistencia de fricción, la cual se debe a que el agua, al igual que los demás fluidos, tiene una cierta viscosidad. Otro es la resistencia de formas, que depende de las líneas hidrodinámicas del casco. Cuando el casco no es de líneas suaves y continuas, los filetes líquidos tienden a apartarse de él, lo que representa un incremento del volumen del agua desplazada. Además, cuando tal separación supera un cierto valor, los filetes líquidos forman unos remolinos y, en tales condiciones, la energía invertida en producirlos se resta a la destinada a mover el buque, constituyendo ésta el componente tercero que interviene en la resistencia.

Los tres componentes citados están íntimamente relacionados entre sí. El cuarto, que se debe a la formación de olas es, por contra, de naturaleza muy distinta; de ahí que se le estudie aisladamente. Al igual que ocurre con los otros, éste aumenta en proporción directa a la velocidad del buque, aunque lo hace de forma muy superior a los demás, por cuyo motivo acaba convirtiéndose en el principal componente de la resistencia a la marcha del buque. Básicamente se calcula por la relación existente entre la eslora del buque y la longitud de las olas que genera en su avance. Sin embargo, para ello es preciso analizar previamente la relación

entre el casco del buque y las olas generadas por él a medida que crece la velocidad.

La roda del buque, al cortar el agua, le transmite una aceleración hacia arriba que varía de acuerdo con la velocidad. Este movimiento vertical actúa en contra de la fuerza de gravedad, tendiendo el agua a ocupar un nivel por encima o por debajo de la superficie, lo que determina la formación de una o más olas que se mantienen estacionarias con respecto al casco, mientras la velocidad de la embarcación permanece constante. Cuando la velocidad es muy reducida, aumenta el número de olas, pues en estas condiciones las moléculas de agua disponen de tiempo suficiente para ascender y descender varias veces a lo largo del paso del buque.

En la roda, el aumento de presión debido a estas olas queda compensado por el descenso de la misma en la popa del casco, la cual tiende a disminuir el nivel de las aguas allí existentes. Estas perturbaciones que aparecen a proa y popa representan una transferencia de energía del buque al agua, lo que conduce a un aumento de la superficie mojada del casco al incrementar el arrastre de agua. Cuando el buque alcanza cierta velocidad, la ola estacionaria de proa está en fase con el seno de la ola de popa, y se acentúan las diferencias de presión en el nivel del agua antedichas; pero si la ola de proa y el seno de popa están completamente desfasados, las presiones se anulan. En el caso de que la velocidad de la embarcación aumente, crecerá la longitud de onda del sistema de perturbaciones; así, un buque acelerado desde cero atraviesa zonas de velocidad progresiva donde la resistencia se incrementa a una tasa muy elevada para ir luego decreciendo. Dicho de otra manera: la curva de resistencia evoluciona de una forma zigzagueante.

Debido a todo ello, el asiento de la embarcación varía. La popa desciende, en tanto que la proa se levanta, y se ve obligada a remontar una colina de agua que ha creado ella misma. Este efecto alcanza un valor crítico cuando la longitud de la onda de proa es igual a la eslora del buque. Entonces, cualquier ligero aumento de velocidad exige un acusado incremento de potencia. De ahí se deduce que, a mayor eslora del buque, más elevada será la velocidad crítica a la que hemos aludido. El casco de mucha eslora y líneas finas reduce considerablemente las perturbaciones que se originan en los extremos de proa y popa, al distribuir mejor a lo largo del casco el empuje que lo mantiene a flote, alejándolo de tales puntos, lo que contribuye a reducir la resistencia de fricción hasta un valor ínfimo en relación con la resistencia total.

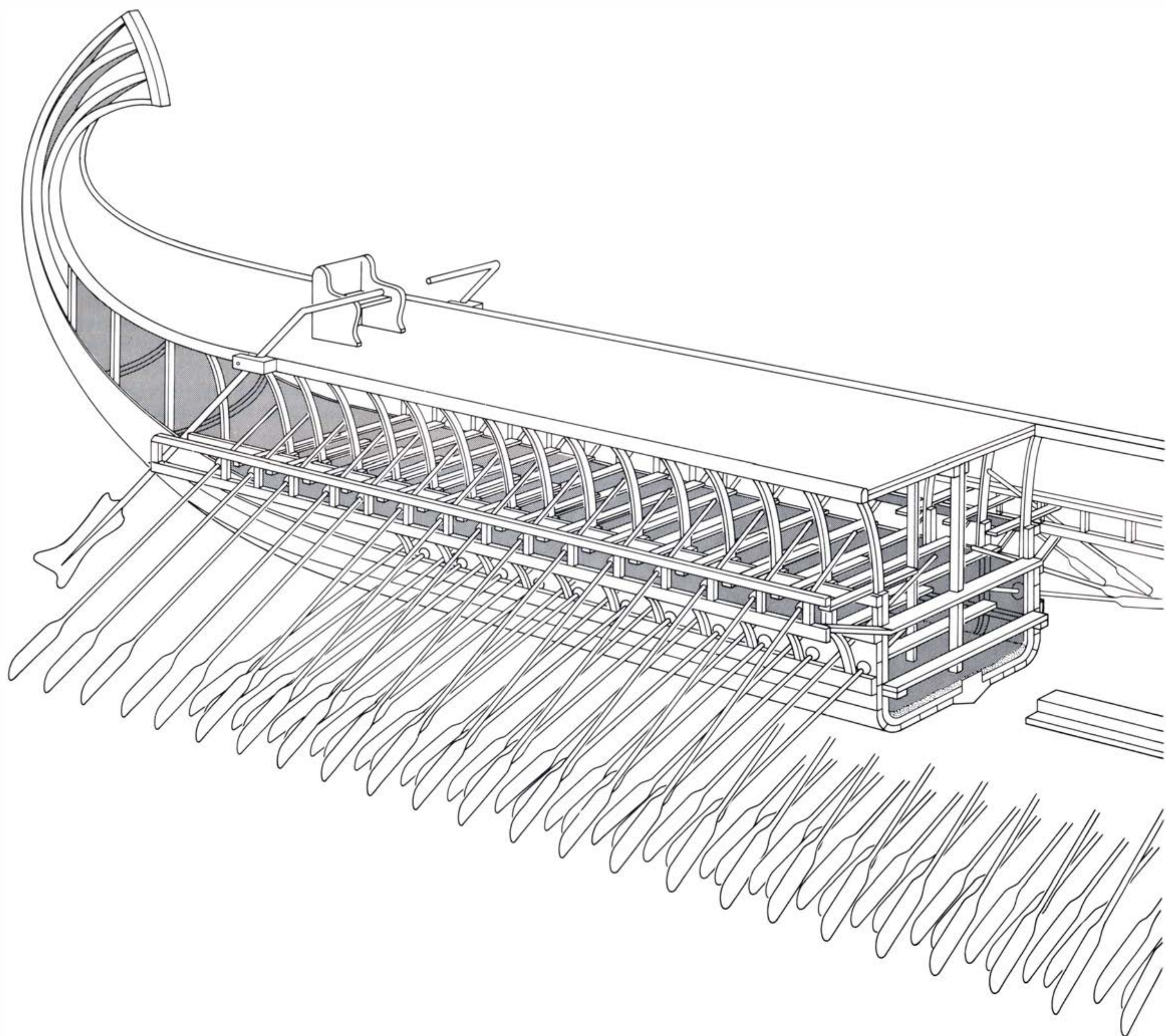
Para una potencia dada, el límite superior de la velocidad de un buque se obtiene por la ley de Froude, concretamente, por la relación entre la eslora y el cuadrado de la velocidad (toda vez que la resistencia aumenta en función del cuadrado de esta última). Esta relación se debe a William Froude, un ingeniero naval británico, el primer científico que clarificó todas estas cuestiones, a mediados del siglo XIX.

Como es lógico, no hay ninguna razón para suponer que los antiguos supieran algo de lo que acabamos de exponer. Sin embargo, por el procedimiento de ir corrigiendo y mejorando



LA MEJOR PRUEBA GRAFICA de un trirreme griego es este fragmento de piedra, esculpido a fines del siglo V a. de C. Conocido como relieve Lenormant en honor de su descubridor, el arqueólogo francés Charles Lenormant, que lo encontró en 1852, se exhibe actualmente en el Museo de la Acrópolis de Atenas. Probablemente está hecho a escala. Al igual que ocurre en la ilustración de la portada, sólo es visible el banco de remeros más alto; los dos bancos restantes se encuentran en el interior del casco o por detrás de los remeros. El trirreme derivó de sencillas embarcaciones de origen griego y fenicio.

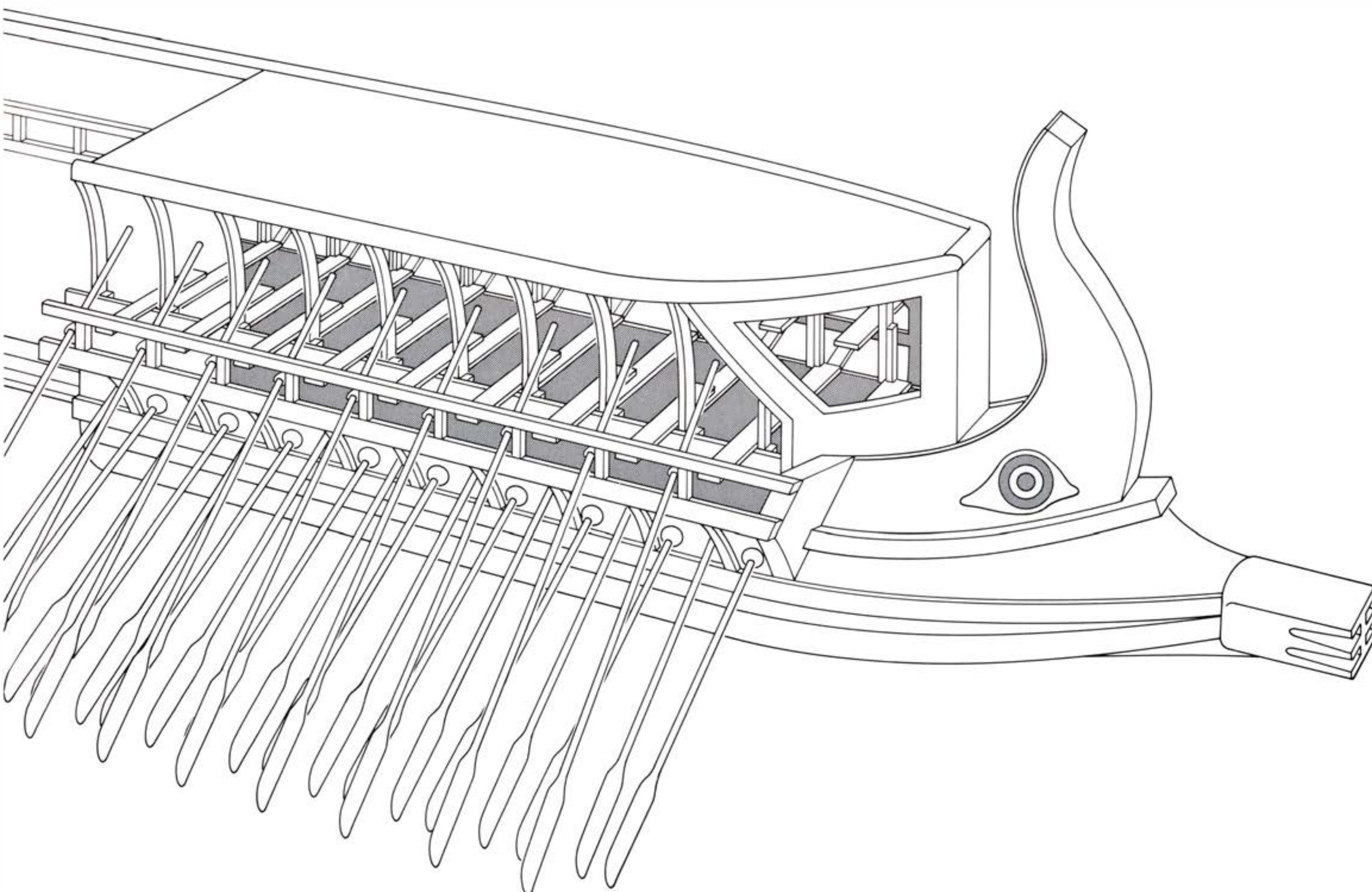
lo existente (método de ensayo y error) llegaron a conocer bastante bien las causas que se oponían al aumento de velocidad de las naves. El casco de las pentecóntoras tenía 38 metros de eslora, y una manga probablemente inferior a 4 metros, lo que representa una relación entre la manga y la eslora de 1:10, la mejor para obtener la máxima velocidad, y que se mantuvo hasta el fin de la era de las naves de guerra propulsadas a remo. Además, todo el mundo coincide en que la eslora de tales naves era aproximadamente la máxima que se podía construir en madera. De hecho, en el caso del trirreme, cuyo casco era aún más fino, todo parece indicar que aquel límite se rebasó ampliamente, gracias al empleo de un complicado sistema de mortajas, mechas y clavijas para unir las tablas del forro, lo que conducía a una perfecta distribución de esfuerzos por todo el casco. A pesar de ello, no eran demasiado seguros: no se botaba al agua ninguno sin antes colocarle un tortor de proa a popa, consistente en un cabo que se tensaba a base de darle vueltas con una palanca o barra de madera. Los puntos de fijación del tortor no se conocen con exactitud,



SECCION DE UN TRIRREME GRIEGO del siglo v a. de C. en la que aparece representada la disposición en escala de los tres bancos ocupados por los remeros, para aprovechar mejor el espacio. Los 170 remeros iban distribuidos de la forma siguiente: 31 en los bancos más altos y, 27, en los intermedios y bajos. A causa de la forma del casco, que se estrecha hacia los extremos de proa y de popa, en aquellas partes sólo había remeros en los bancos altos.

pero sí se sabe que lo ponían para apretar las costuras y evitar que hicieran agua al verse sometido el casco a cualquier esfuerzo. La madera es muy reacia a mantener la estanqueidad cuando está sometida a tensión.

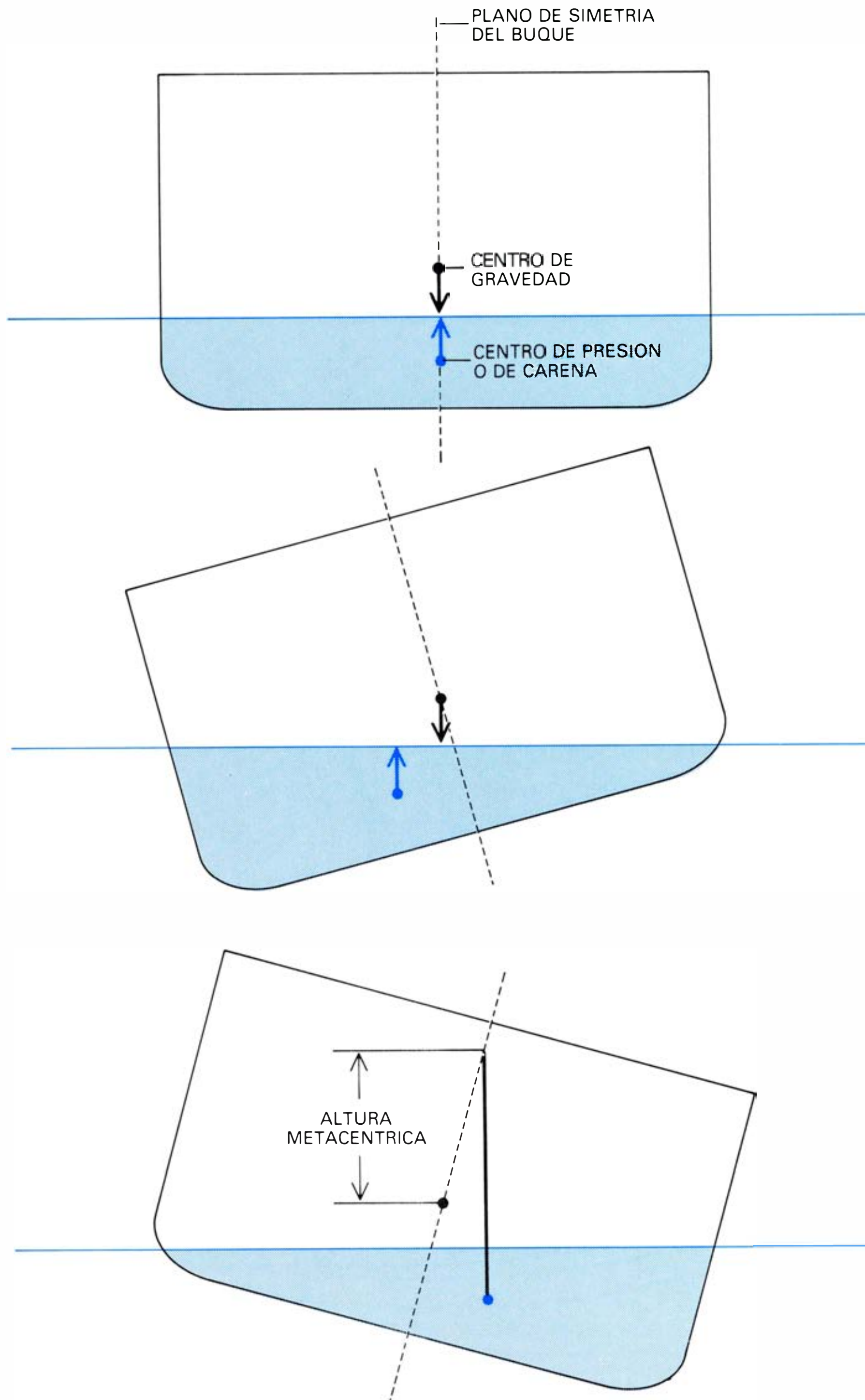
La manga de aquellas naves o galeras era la mínima indispensable para que dos remeros, sentados de lado, tuvieran espacio suficiente para manejar los remos, que se apoyaban en la regala de la embarcación, es decir, en la tabla superior del forro del costado. El fulcro o punto de apoyo del remo debe estar siempre bastante alejado del puño del mismo, pues de lo contrario el manejo exige una fuerza excesiva. La norma actual consiste en colocar un tercio o algo menos de la longitud del remo dentro de la embarcación. Por otro lado, teniendo en cuenta que la longitud de los remos empleados en los trirremes era muy semejante a los actuales, cabe suponer que en la antigüedad utilizaran la misma proporción que acabamos de indicar. Además, si tenemos en cuenta que debe existir una cierta separación entre los puños o extremos interiores de los remos para que quepan los remeros, se deduce que en las pente-



La eslora total de un trirreme era, aproximadamente, de 35 metros, con 3,5 metros de manga. El espolón era su principal elemento ofensivo. El gobierno se hacía mediante dos timones de caja situados en las aletas. Esta representación se basa fundamentalmente en las investigaciones realizadas por el británico J. S. Morrison, el cual, a su vez, lo hace en las proporciones del relieve Lenormant. Un trirreme así debió alcanzar una velocidad máxima de 11,5 nudos.

cóntoras éstos iban muy apretados y dejando poco espacio libre entre ellos. El calado de estas naves, muy pequeño, oscilaba posiblemente alrededor del medio metro. De ahí que tuvieran un desplazamiento y una resistencia de fricción muy reducidos.

El calado de tales galeras era muy pequeño, principalmente porque estaban hechas de madera muy delgada y ligera. Las tablas del forro exterior medían unos 3,5 centímetros de grueso, y en algunas partes se hacían aún más delgadas. Al referirse a las naves mercantes, normalmente de construcción más robusta, los poetas de la época señalaban que sólo una madera de tres dedos de grueso separaba al marinero de su perdición. La delgadez de las maderas llegaba hasta el extremo de que la tripulación pesaba una tercera parte del conjunto. Por todo ello podemos calcular que el peso total de un trirreme, incluyendo los remeros, era inferior a las cuarenta toneladas métricas.



ARQUIMEDES, que vivió en el siglo III a. de C., fue el primero en investigar de forma sistemática la estabilidad del buque. Supuso que el peso del mismo se concentraba en el centro de gravedad y apuntaba hacia abajo, en dirección al centro de la tierra (*flecha negra*). El empuje, dirigido hacia arriba (*flecha de color*), viene aplicado en el centro de masa, o de carena, que es el centro del volumen sumergido (*arriba*). Cuando la nave escora, las dos fuerzas citadas forman un par que tiende a hacerle recobrar la vertical, por cuanto el centro del volumen sumergido se ha desplazado hacia el lado donde ha escorado (*en medio*). En términos náuticos actuales se dice que el centro del impulso de Arquímedes se halla a cierta distancia vertical de la línea media del buque; la distancia vertical existente entre el centro de gravedad y la intersección de la vertical que pasa por el citado centro de empuje con el plano de simetría del buque se denomina altura metacéntrica (*abajo*). Cuanto mayor sea esta altura, tanto más estable será el buque.

Las características del casco favorecían el que las naves avanzaran a gran velocidad. Las líneas de los petroleros actuales son muy parecidas a las de aquellas, debido a la proa de bulbo que se proyecta hacia delante por debajo del agua, con el fin de impedir la formación de la ola de marcha, por cuanto evita que el agua experimente acusadas variaciones de momento. Los espolones de la antigüedad tenían una forma algo distinta, aunque probablemente producían unos efectos parecidos. La popa de las galeras se elevaba del agua de manera gradual y progresiva, como ocurre en cualquier yate de regatas, lo que evitaba la formación de remolinos y estela. De todos modos, las galeras antiguas se diferenciaban perfectamente de las embarcaciones modernas: éstas suelen llevar una quilla de aleta en la que va articulado el timón. Las naves antiguas, por el contrario, se gobernaban por medio de unos remos, espadillas o timones de caja montados en las aletas, los cuales impedían la formación de remolinos, a la vez que reducían al mínimo el área del casco mojada.

Por encima del agua, aquellas galeras respondían igualmente a un proyecto muy depurado: el reducido puntal hacía que la resistencia que ofrecían al viento fuera muy pequeña, salvo en la popa, donde la curva del casco se prolongaba hacia arriba hasta formar una especie de cresta o abanico. Este remate protegía el casco no sólo de la posibilidad de entrada de golpes de mar cuando navegaba en popa, sino también, y como consecuencia de aumentar el momento de inercia de giro del mismo, reducía el balance yendo en lastre. Dicha cresta servía igualmente para producir un efecto de orza en la galera cuando incidía sobre ella una racha de viento, lo que minimizaba el posible riesgo de que una ola rompiera contra el costado y la inundara.

Mucho antes de la aparición del trirreme, la galera había alcanzado ya unas cualidades marineras muy notables; en lo relativo a velocidad, las pentecóntoras más rápidas andaban

unos 9,5 nudos, o sea, unos 17,6 kilómetros por hora, como máximo, lo que viene a ser algo así como un nudo menos que las embarcaciones de regatas modernas.

Además de sus características funcionales, las galeras llevaban el marchamo particular y propio de cada constructor naval. La estanqueidad del casco se obtenía a base de unas manos de brea y, en ambas amuras, más arriba del espolón, había unos ventanillos para la ventilación del interior o unos puntos pintados de forma muy destacada, que evolucionaron hasta convertirse en los ojos que se observan aún hoy en algunas embarcaciones. Normalmente, el espolón era de bronce y terminaba en forma de varias puntas de flecha o de morro de un gigantesco jabalí. Por todo ello, cuando la galera se aproximaba a otra nave para embestirla con el espolón, resultaba muy llamativa. De color negro y con la peculiar cresta levantada en la popa, mostraba la imagen feroz de una bestia salvaje. A veces colgaban de sus costados una especie de pavesas de cuero teñido y sin quitarle el pelo. Los remos contribuían aún más a darle un aspecto zoomórfico: Los poetas de la época comparaban frecuentemente el movimiento sincronizado de las palas con el de las alas de los pájaros. A veces, las naves presentaban incluso una irisación propia de los animales. Aristóteles relataba cómo la espuma que levantaban los remos al meterlos en el agua producía el arco iris cuando los rayos solares incidían convenientemente en ella.

De hecho, los constructores de galeras se preocupaban fundamentalmente por el aspecto funcional y, desde los tiempos de Homero hasta el año 500 a. de C., aproximadamente, aparecieron varias mejoras tendentes a aumentar la potencia motriz de la pentecóntora. No conocemos con exactitud todos los detalles de estas mejoras; además, la historia pormenorizada de esta evolución resulta excesivamente compleja para resumirla aquí. Consistió, en esencia, en aumentar la altura del casco, añadiéndole varias cubiertas o bancos donde poder colocar un mayor número de remeros. Así, la colocación de una segunda fila de remeros dio lugar al birreme o galera de dos órdenes de remos y, al trirreme, la adición de una tercera.

El logro de todas estas mejoras, sin causar ninguna reducción en la estabilidad de la galera, de por sí muy escasa, obligó a proceder con mucho cuidado. En el caso de un birreme, de haber colocado la segunda fila u orden de remeros más arriba de las cabezas de la otra, la galera hubiera volcado fácilmente por exceso de peso alto. Por tal motivo la solución fue elevar ligeramente los costados de la nave, sin que en ningún caso el aumento de altura representara más de medio metro, y colocando los remeros de la fila de arriba de modo que cada uno de ellos estuviera sentado en medio de dos de la otra, situados más bajos, en lo que podríamos llamar la bodega de la nave. Los primeros apoyaban los remos en la regala, como antes, mientras los segundos, que se encontraban a escaso medio metro de la superficie del agua y no era aconsejable, por tanto, rebajar dicha regala, usaban unos remos que salían por unos agujeros o portas situados en los costados. Tales agujeros iban debidamente protegidos con unos pedazos de piel, para impedir la posible entrada del agua.

El paso de birreme a trirreme fue viable gracias a la colocación de una pieza nueva, la postiza, situada a la altura de la regala, mas por el lado de fuera; ello permitió a los constructores navales alejar el fulcro o punto de apoyo de los remos unos sesenta centímetros de la borda. Los remeros que manejaban este tercer orden de remos iban sentados a la altura de los hombros de la gente que ocupaba la fila intermedia, y no por encima de sus cabezas, de modo que nuevamente el aumento del puntal de las naves no superó los 50 centímetros.

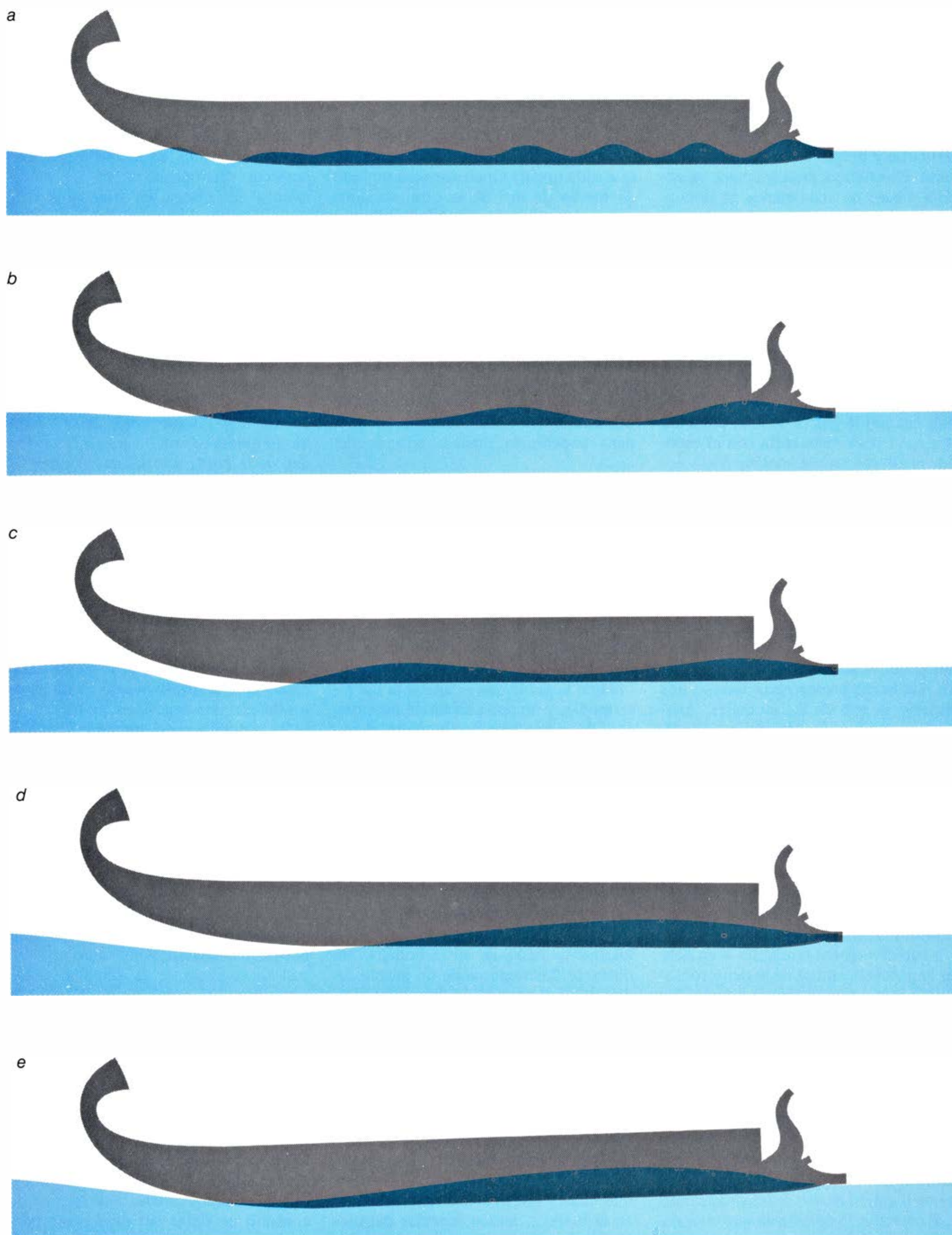
Algo más adelante, la estrechez de los remeros disminuyó al colocarlos en escala, es decir, situando los hombres de la fila más alta algo así como medio metro por delante de los que estaban en los bancos intermedios, y éstos, a su vez, lo mismo con respecto a los que se hallaban en la bodega. No obstante, los remeros de un mismo orden o fila iban muy juntos: la separación entre ellos era de un metro más o menos. Esto significaba que, al bogar, si alguno lo hacía ligeramente fuera de compás recibía en la espalda el golpe que le propinaba con los nudillos o con el puño del remo el que estaba sentado detrás de él o, por el contrario, era él quien golpeaba al situado delante suyo.

Los hombres que ocupaban el orden más bajo, y que por tanto iban sentados en la bodega, tenían, además del problema de espacio, otro de índole muy distinta. Como señaló el poeta cómico Aristófanes en la obra *Las ranas,* al remar e inclinarse hacia delante aproximaban excesivamente las narices a las posaderas de los pertenecientes a la fila intermedia, en el preciso instante en que éstos se inclinaban también hacia delante. En estas condiciones, y al hacer el esfuerzo propio de la boga, era muy frecuente que se les escapara algún viento inoportuno.

La existencia de tres órdenes de remeros dispuestos en escala exigía una perfecta sincronización de los movimientos de la boga. Es decir, el paralelismo que se observa en los remos en cualquier representación longitudinal de una nave de ese tipo debía mantenerse constantemente y sin que ninguno de ellos se apartara más de 30 centímetros de aquella condición, pues de lo contrario chocarían entre sí. Y en el caso de que esto ocurriera, se produciría un fenómeno similar al que ocurre con las fichas de dominó, propagándose el choque a todos los remos de una banda, lo que afectaría no sólo a la velocidad de la nave, sino también al rumbo de la misma.

Para lograr que los remeros bogaran al compás se realizaban ejercicios constantemente, al mismo tiempo que se procuraba darle a la gente unos buenos incentivos. En la antigüedad, los remeros pertenecían a la categoría de hombres libres, los cuales tenían lógicamente un gran interés por la supervivencia de su ciudad-estado. Por lo general, los esclavos se empleaban sólo en momentos de extrema necesidad, en cuyo caso se les concedía previamente la libertad. Lógicamente bastaba un sólo hombre descontento a bordo para entorpecer y arruinar la boga de los demás. El látigo no se usó jamás. Cada nave llevaba un salomador que, a base de cantos, se encargaba de mantener el ritmo de la boga; los remeros percibían un salario bastante alto, lo que representaba un buen incentivo. De todos modos, normalmente procedían del estrato más pobre de la ciudad, y por tanto no estaban en condiciones de adquirir las armas propias de la infantería. Sea por la razón que fuere, lo cierto es que ponían siempre un gran empeño en la boga.

Entre los remeros del trirreme, siempre empapados en sudor, y los marineros armados que iban en cubierta listos para entrar en acción tan pronto les fuera posible, debía existir lógicamente una cierta tensión. En los estados más democráticos, cual Atenas, la cuestión no debía ser muy grave, por cuanto toda la estrategia se basaba en el espolón, y en consecuencia el número de gente armada a bordo estaba reducido al mínimo indispensable. Los trirremes atenienses que intervinieron en la guerra del Peloponeso llevaban



OLEAJE que produce en el agua el paso del buque. Constituye la expresión más importante de la resistencia del fluido al avance del mismo. A velocidad reducida (*a*), la formación de las olas afecta escasamente al buque. Pero a medida que ésta aumenta, lo hacen también la longitud y la altura de las olas. La ola de proa puede disminuir o aumentar la de popa, según ambas estén desfasadas (*b*) o en fase (*c*). Finalmente, la cresta de la ola de proa tiende a levantar la cabeza del buque, al mismo tiempo que el seno de la que se encuentra en el extremo opuesto hace descender la popa (*d*). Cuando esto sucede, el asiento del buque varía, y ello produce que una parte cada vez mayor de potencia se invierta en hacerlo remontar la ola (*e*). Este inconveniente se evita haciendo la eslora del buque tan grande como sea posible. Por cuestión de claridad, las olas se han exagerado notablemente en la ilustración.

tan sólo 14 marineros. Los remeros iban perfectamente distribuidos a bordo, de modo que 170 de ellos cabían en una nave de aquel tipo, de dimensiones muy parecidas a la pentecóntora.

El resultado de todo ello era que los trirremes desarrollaban una velocidad realmente alta y al mismo tiempo ostentaban una excelente capacidad de maniobra. Unos cálculos muy razonables permiten deducir que la máxima de un trirreme era de 11,5 nudos, unos 21,3 kilómetros por hora. Sin embargo, tales cálculos son en cierto modo muy conservadores, por cuanto se supone que el casco corta y desaloja el agua a medida que avanza, pese a la opinión de algunos proyectistas navales que consideran que los trirremes, al ser muy ligeros y veloces, podían planear en el agua, al igual que hacen hoy muchas embarcaciones deportivas, y en cuyo caso la velocidad máxima podría aumentar en un 50 por ciento. Sin embargo, el ritmo de la boga para desarrollar la velocidad máxima dejaba rápidamente agotados a los remeros, y por tal motivo lo podían mantener sólo durante cinco o diez minutos. En cualquier caso, la velocidad máxima del trirreme era comparable a la carga de caballería medieval.

Los trirremes necesitaban unos treinta segundos para alcanzar la velocidad máxima, partiendo del reposo; en cambio, la mitad de esa velocidad la conseguían en sólo ocho segundos y la cuarta parte en dos. Debido a esto no necesitaban mucho tiempo para coger arrancada y poder embestir con el espolón, asestando un golpe de cierta contundencia. Además, los trirremes eran excelentes para arrojar dardos. También maniobraban con gran facilidad haciendo la ciaboga, o sea, remando hacia atrás la gente de una banda y al revés los de la otra, de modo que podían revirar la nave en un espacio ligeramente superior a una eslora. El espectáculo de contemplar una flota de unas cuantas docenas de naves de este tipo navegando o dirigiéndose hacia el puerto de procedencia después de realizar unas maniobras debía ser verdaderamente impresionante.

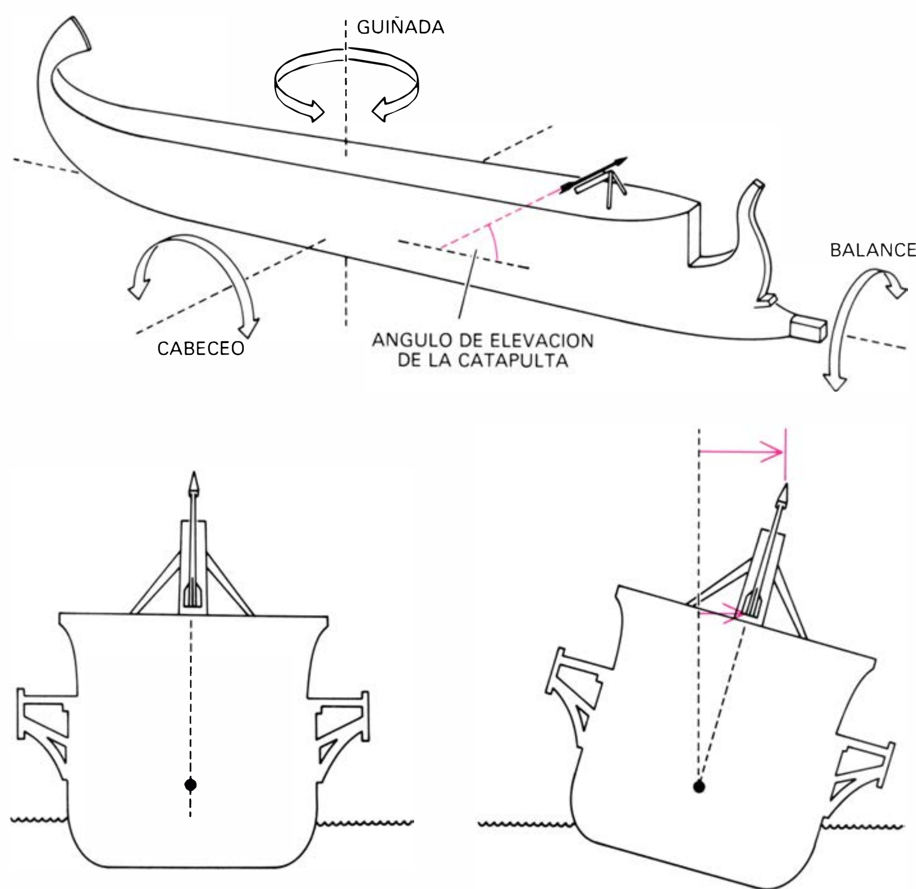
Uno de los principales viajes rápidos efectuado por los trirremes a un lugar situado a mucha distancia ocurrió en el 427 a. de C. En aquella ocasión, la ciudad de Mitilene, en la isla de Lesbos, se había levantado contra los atenienses y éstos sofocaron la rebelión. El demagogo ateniense Cleón propuso como castigo el exterminio de toda la población. Gracias a sus dotes de orador consiguió que la asamblea lo aprobara. Inmediatamente, un trirreme se hizo a la mar con el fin de llevar la orden a la guarnición ateniense. Sin embargo, debido a la gran efervescencia política existente en Atenas, la nave no debió salir hasta primeras horas de la tarde, poco después de haber terminado la votación. Y como dice Tucídides, dado el macabro encargo del cual era portadora, no debía navegar con mucha prisa; por tal motivo, remando sólo uno o dos órdenes de remos y dando estrepadas cortas, andaría algo así como a cuatro o cinco nudos. A la mañana siguiente, cuando la asamblea se reunió de nuevo, prevaleció el sentido común, lo que significaba la anulación de la orden de efectuar aquella masacre. Los embajadores de Mitilene en Atenas habían previsto la posibilidad de que se produjera este cambio de opinión y habían preparado una galera rápida, aprovisionada con vituallas de alto valor energético y tripulada por gente seleccionada, a la que prometieron una buena suma de dinero si conseguían alcanzar la nave salida el día antes.

Esta segunda nave debió de partir unas 24 horas más tarde hacia Lesbos, a unos 345 kilómetros de distancia. Cuando llegó a mar abierto había oscurecido ya y la gente estuvo remando durante toda la noche, comiendo de vez en cuando galletas de cebada mojadas con vino. La noche era lo suficiente clara para la navegación y por suerte no había viento de proa. Con el fin de mantener la velocidad máxima, no se sabe exactamente si los embajadores contrataron el número de gente necesaria para poder relevar a todo un orden de remeros y establecer así unos turnos de boga para mantener en todo momento a uno descansando, o por el contrario bogaban sólo dos órdenes mientras el tercero descansaba. Independientemente de cuál de estos dos sistemas utilizaron, lo cierto es que llegaron a Mitilene a mediodía, inmediatamente después de la otra nave. Todo esto parece indicar que invirtieron en la travesía menos de veinticuatro horas, lo que representa un andar de casi nueve nudos, unos 16,6 kilómetros por hora. Cuando llegaron, la orden había sido dada, pero afortunadamente no habían tenido tiempo todavía de cumplirla. Los transbordadores actuales hacen la misma travesía en 14 horas.

El siguiente progreso de importancia que afectó a las galeras ocurrió hacia el 400 a. de C. Por aquel entonces, iniciaron su aparición las naves de cuatro órdenes, conocidas por cuadrirremes, y el año 399 un conjunto de constructores reunidos por Dionisio de Siracusa hizo el primer quinquerreme, o galera de cinco filas u órdenes de remeros. En ambos casos, aunque todo parece indicar que se trataba de una continuación del sistema establecido por las galeras de un sólo banco u orden de remos, los birremes y los trirremes, en realidad no es así. Todos los estudios realizados al respecto indican que jamás se construyeron galeras con más de tres filas de remos superpuestas. Desde el punto de vista material esta opinión se justifica por cuanto el manejo de los remos situados en la cuarta fila u orden, y debido al ángulo de trabajo de los mismos, hubiera exigido un esfuerzo excesivo. Aun en el caso de un trirreme, el que realizaban los remeros del banco situado más arriba era considerablemente superior al de los demás, de ahí que quienes lo ocupaban recibían normalmente una paga algo más elevada.

Al parecer, en vez de aumentar el número de niveles de bancos, lo que hicieron fue colocar dos hombres en cada uno de ellos. De acuerdo con la forma del casco del trirreme, la colocación de un hombre más en cada banco empezó en el orden situado más arriba, convirtiéndose el trirreme en cuatrirreme. Sólo se podía instalar allí una fila más de remeros sin alterar demasiado la estructura de la nave. Los cartagineses, por el contrario, adoptaron verosímilmente una solución alternativa. Las galeras utilizadas por éstos se parecían mucho a las fenicias y tenían más manga; de modo que podían acomodar probablemente a cuatro remeros, sentados uno al lado del otro, y situados a un mismo nivel. Sin embargo, todo parece indicar que la evolución de la galera partió de las naves de poca manga.

El paso del cuadrirreme al quinquerreme obligó a introducir modificaciones en el casco, a menos de suponer que el nuevo orden de remeros iba pegado al fulcro, en cuyo caso la potencia producida por éstos habría sido prácticamente nula. Por esta razón, lo más lógico es suponer que los constructores optaron por bajar la postiza y simultáneamente poner otra similar más arriba, lo que permitía colocar dos remeros en cada uno de los bancos altos e intermedios. Los bancos situados más abajo siguieron probablemente con un sólo hombre, con el fin de conservar la velocidad que proporcionan los cascos de obra viva muy fina, aun cuando el calado de los cuadrirremes y quinquerremes había aumentado por efecto del



PODEMOS DIVIDIR LOS MOVIMIENTOS del buque en guiñada, cabeceo y balance (*arriba*). Los movimientos tienen una gran incidencia en la puntería de un arma cual la catapulta. En el caso de las galeras antiguas, la notable proporción entre la manga y la eslora hacía que fueran 50 veces más sensibles al balance que al cabeceo. Al aumentar el ángulo que forma la catapulta con la horizontal para dar más alcance al proyectil, la cabeza del mismo se alejaba más del centro de gravedad de la nave que su pie o extremo posterior (*abajo, a la izquierda*). En estas condiciones, cuando la nave balanceaba, la cabeza del proyectil se desplazaba más hacia la banda que no el pie del mismo (*abajo, a la derecha*). La puntería se podía hacer bien únicamente cuando el proyectil estaba en un plano vertical, en cuyo caso la probabilidad de dar en el blanco era elevada. En cambio, si se disparaba en el instante en que la nave iniciaba el balance, cuando estaba ligeramente fuera de la vertical, se desviaba hacia uno u otro lado del blanco. La eficacia de los disparos de la catapulta disminuía considerablemente cuando se hacían más o menos hacia proa; por ejemplo, si se disparaba contra un blanco que se pretendía embestir con el espolón.

peso de los nuevos órdenes de remeros colocados a bordo. En los bancos altos de los trirremes iban originariamente 31 remeros por banda, mientras que los dos restantes llevaban 27 cada uno. Por todo ello los cuadrirremes recién aparecidos debieron contar con un total de 232 remeros, en tanto que en los quinquerremes la cifra se elevaría a 286.

Los primeros documentos gráficos que ilustran estos cambios son de un siglo posterior a la aparición de los cuadrirremes y quinquerremes. De acuerdo con las exigencias prácticas de la boga, tales ilustraciones demuestran que en las galeras griegas de mayor tamaño, de los siglos tercero y segundo antes de Cristo, todos o gran parte de los remos se apoyan en una postiza de notable longitud, situada no en la regala sino a cierta distancia de ella. Los romanos siguieron la tradición cartaginesa y prefirieron los cascos de mucha manga y con portas para los remos. De hecho, no siendo buenos marinos, preferían abordar las naves enemigas y luego invadirlas con los soldados, convirtiendo las batallas navales en unos combates con tácticas de infantería, en los que eran unos excelentes maestros. En este caso, es lógico que a los romanos no les preocupara demasiado la velocidad de las naves, sino que las preferían con mucho espacio para poder embarcar un gran contingente de tropa.

Durante las primeras décadas, los cuadrirremes y quinquerremes se difundieron de forma muy lenta en todas las marinas del Mediterráneo. El 330 a. de C. la flota de Atenas contaba sólo con 18 cuadrirremes, frente a 492 trirremes. Seis años después, aquel número había aumentado a más del doble, al mismo tiempo que había empezado la construcción de quinquerremes. Para conocer las causas de esta lentitud inicial es preciso comparar las principales características físicas de los trirremes, cuadrirremes y quinquerremes.

En lo que respecta al trirreme, uno de los principales problemas que debían solucionar los constructores era la necesidad de evitar los pesos altos. Las condiciones que hacen un buque estable y tenga, por consiguiente, la propiedad de adrizarse cuando escore hacia una u otra banda, se explican hoy en día por la altura metacéntrica del mismo. Este concepto, sin embargo, no aparece hasta el siglo XVIII, aunque la mayor parte de conocimientos físicos y matemáticos que intervienen los conocía ya Arquímedes. Según se desprende de algunos indicios, Arquímedes estudió los cuerpos flotantes por algo relacionado con la estabilidad de las naves. Si consideramos la sección transversal de una nave típica [*véase la ilustración de la página 108*], su centro de gravedad estará en algún punto situado en el interior de la misma y que por razón de simetría se encontrará en el plano diametral.

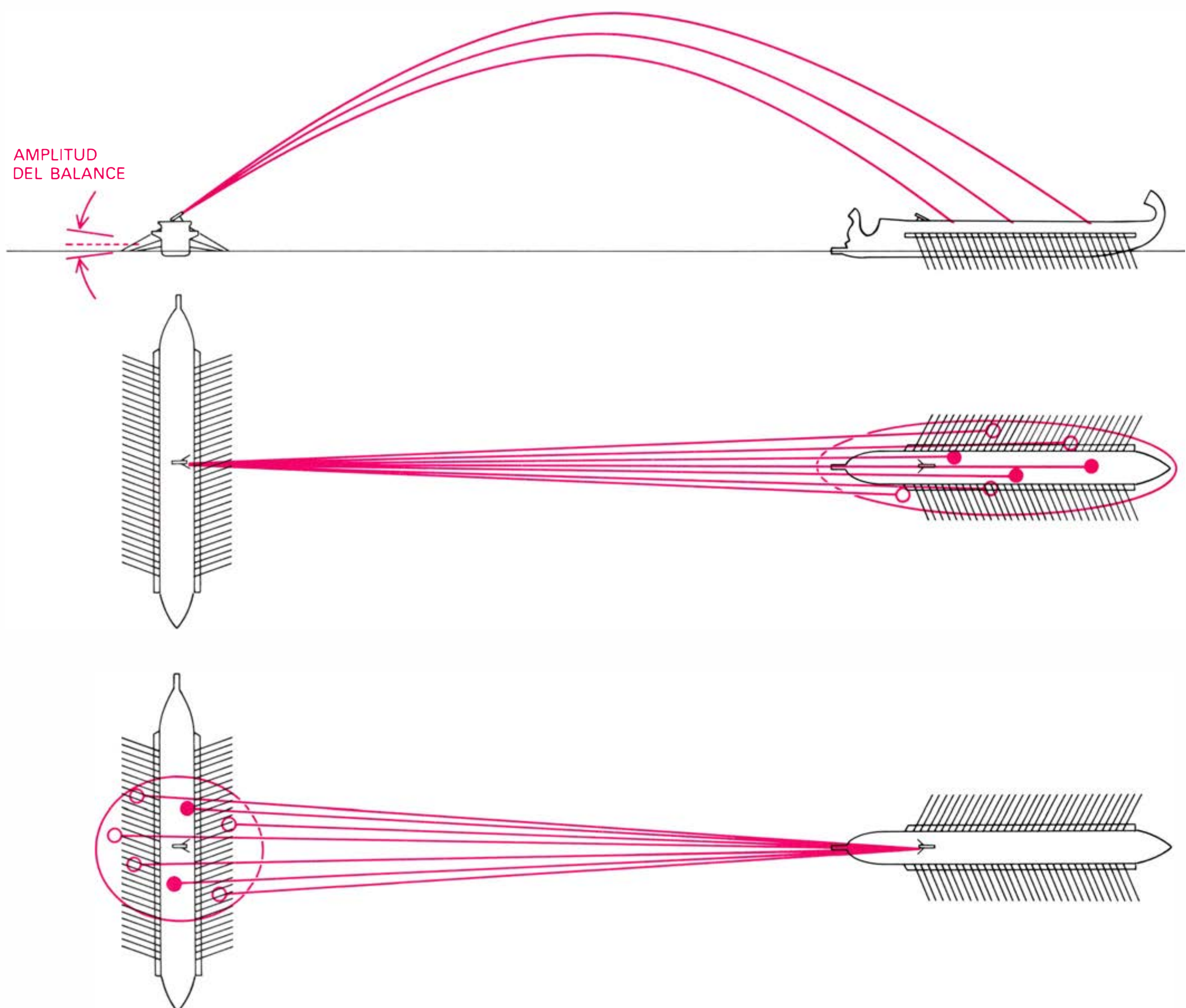
Por otro lado, podemos imaginar que la nave se comportará como si su masa estuviera concentrada en dicho punto y tendiera a dirigirse hacia el centro de la tierra en virtud de la gravedad. En estas condiciones, si la nave escora o se inclina hacia una banda, el centro de carena, llamado también centro de presión y que en realidad es el centro de volumen de la parte sumergida del casco, se desplaza paralelamente a la línea que une los centros de gravedad de las cuñas de emersión e inmersión, en dirección hacia este último. En cualquier caso, el empuje del agua viene aplicado en el centro de carena y se dirige hacia arriba. Cuando el buque escora, dicho empuje forma con el peso del buque un par de fuerzas que tienden a adrizarlo. La altura metacéntrica a que antes hemos aludido es sencillamente la distancia existente entre el centro de masa del buque y el punto donde el empuje de Arquímedes en sentido ascendente corta el plano diametral de la embarcación. De acuerdo con todo esto, y según es posible comprobar en la ilustración, cuanto mayor sea la altura metacéntrica más estable será el buque.

Sabemos con bastante exactitud las dimensiones de los trirremes, así como el lugar que ocupaban los tripulantes y demás efectos de a bordo, lo que nos permite deducir la altura metacéntrica. Conviene advertir, no obstante, que los valores obtenidos son susceptibles de ligeras variaciones por razón del calado, el peso de la madera del casco y la distribución de los remeros a

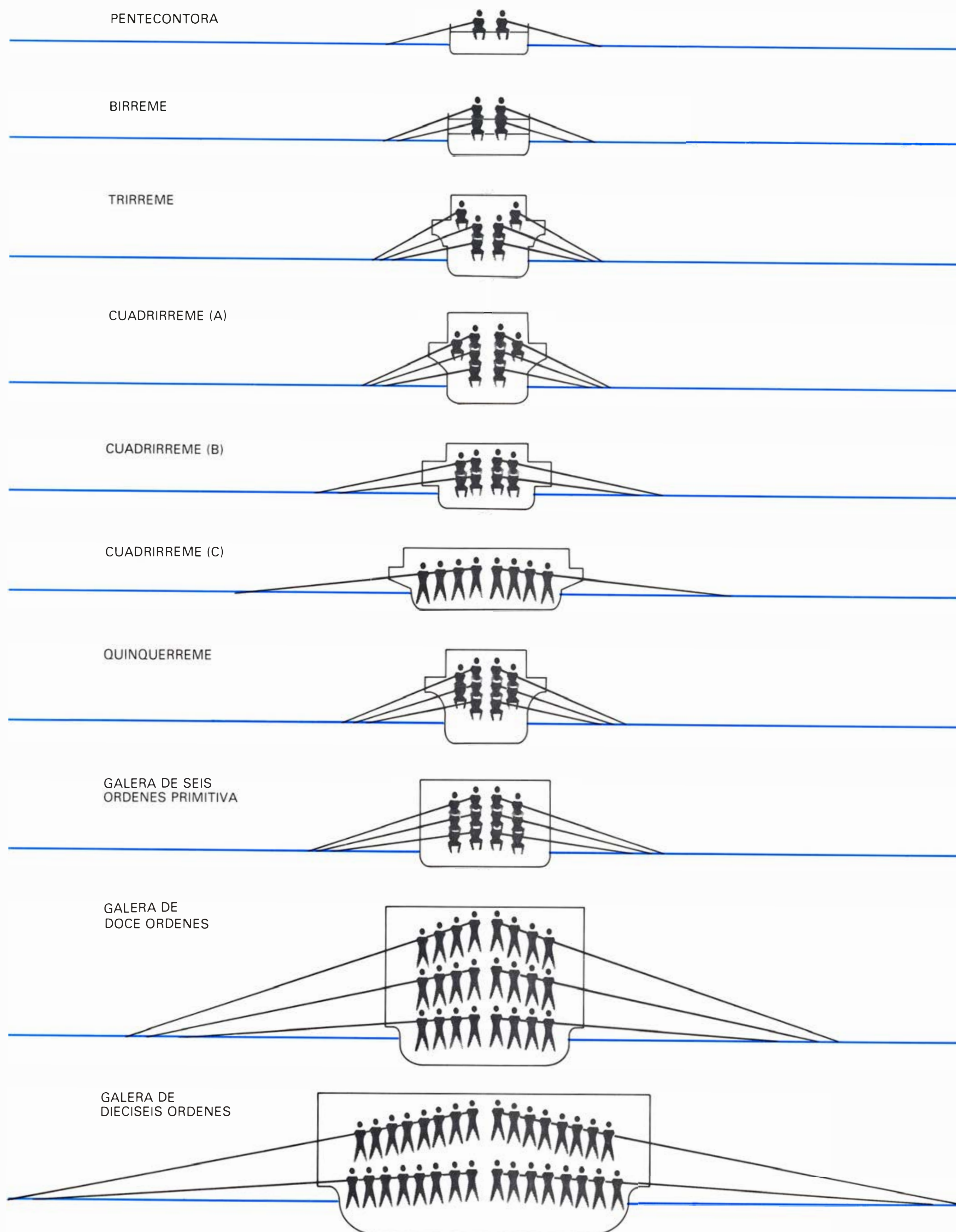
bordo. Sin embargo, cualquier ligera variación de estos factores puede producir sólo unos cambios del orden de cinco o seis kilómetros por hora en la intensidad del viento capaz de tumbar la nave. Así, empleando unos parámetros constantes para los diversos tipos de naves, la relación de valores metacéntricos que obtengamos será totalmente válida aun cuando ninguno de los números utilizados sea rigurosamente exacto. Para que esto ocurra es preciso suponer que, al producirse el paso del trirreme al cuadrirreme o quinquerreme, tanto la estructura del casco como todo lo demás sufrieron las modificaciones mínimas necesarias. Esta hipótesis resulta muy aceptable, si tenemos en cuenta que los constructores navales se han mostrado siempre muy reacios a introducir cualquier cambio.

En este análisis el principal imponderable es el lastre, por cuanto en el interior de las galeras había espacio suficiente donde colocarlo y alterar la estabilidad de las mismas. Ha de aclararse que en el cómputo hemos procurado que el peso del conjunto fuera lo más ligero posible, con el fin de mantener la idea de los constructores navales de la época en el sentido de que la nave respondiera inmediatamente al menor esfuerzo de los remos. De acuerdo con este criterio hemos supuesto que dicho lastre era de 13.000 kilogramos, cantidad que consideramos suficiente para evitar que los vientos de 60 o menos kilómetros por hora la tumbaran.

Así, llegamos a la conclusión de que la altura metacéntrica de los trirremes era de 0,4 metros, la cual, aún siendo muy adecuada, no la podemos calificar de exagerada. Este valor sirve para justificar plenamente la precaución de los antiguos, en el sentido de evitar los combates navales cuando la mar estaba agitada. En el caso del quinquerreme, la adición de cuatro órdenes más de re-



EFICACIA DEL FUEGO DE CATAPULTA disparado desde una galera que estuviera a la defensiva, o sea, que intentara evitar la embestida de un espolón; dicha eficacia se veía muy favorecida por las características de su propia estabilidad y por la forma de la galera atacante que constituía el blanco. Si se tiene en cuenta que tales naves eran muy sensibles al balance, los proyectiles disparados por la nave a la defensiva, en dirección prácticamente perpendicular al plano diametral de la misma, experimentarían una dispersión en distancia, haciendo blanco en una zona de forma parecida a una elipse alargada (*arriba y centro*), que coincide bastante aproximadamente con la figura de la galera atacante, por cuyo motivo la probabilidad de hacer blanco era muy grande. (Al mismo tiempo, la disposición de los bancos en escala aumentaba la posibilidad de que cada impacto lesionara a más de un remero.) Por el contrario, los proyectiles lanzados por la nave atacante, y como consecuencia del balanceo, experimentaban una notable dispersión en amplitud, por cuanto dispararía más o menos hacia proa (*abajo*). En estas condiciones, la citada dispersión no sería ningún inconveniente, aunque, debido a la poca manga del blanco, cualquier ligero cambio en la elevación de la catapulta resultaba suficiente para que los proyectiles no llegaran o pasaran de largo del blanco. De todo esto se desprende que el fuego de catapulta era mucho más efectivo para la nave que estaba a la defensiva que para la embarcación atacante.



EVOLUCION DE LAS GALERAS de la antigüedad, se produjo en tres fases fundamentales. La primera, ocurrida hacia el 800 a. de C., coincidió con la invención del espolón, y condujo a la sustitución de las naves de un solo orden, o nivel de remos, tales como los pentecóntoras de 50 remos (*arriba*), por los birremes y luego, tras la aparición de las postizas, por los trirremes. Con el trirreme culminó la primera fase de la evolución de las galeras, por cuanto era imposible aumentar los tres niveles o alturas que ocupaban los remeros. La segunda fase consistió en incrementar el número de remeros que bogaban en cada banco y remo. Esto condujo a una gran diversidad de soluciones por cuanto existen varias formas de combinar los remeros y bancos. De todos modos había un límite, pues el número de remeros por banco y remo no podía ser nunca superior a ocho (*abajo*). En este caso los remeros bogaban de pie y andaban hacia delante y hacia atrás para dar las estrepadas. El estadio final del proceso evolutivo se dio con la construcción de catamaranes gigantescos.

meros en las dos filas más altas de a bordo origina un dilema al contraponer la velocidad de la nave con la estabilidad. Con una obra viva igual a la del trirreme, e idénticos supuestos que los establecidos antes, el cálculo de la altura metacéntrica del quinquerreme da 0,1 metros. En estas condiciones, los balances serían muy lentos y bastarían vientos de algo más de 30 kilómetros por hora para tumbarlo.

Este resultado obliga a suponer que los constructores debieron introducir en las naves alguna modificación de tipo estructural. La única solución válida era aumentar la manga, lo que significaba una reducción de la velocidad, cuando la razón principal que impulsó a aumentar el número de órdenes fue incrementada. En el quinquerreme, si el aumento de la manga se hizo de modo que la estabilidad resultante fuera igual a la del trirreme, se conseguía sólo un aumento de velocidad de un 14 por ciento, con respecto a éste, en tanto que su aceleración sería mucho más reducida. Unos resultados francamente pobres, habida cuenta del aumento del número de tripulantes, que suponiendo fuera de 62 remeros en el orden más alto y de 54 en el intermedio, venía a oscilar alrededor del 70 por ciento. De todos modos, si admitimos para el quinquerreme una estabilidad tan reducida como hemos obtenido anteriormente, el aumento de velocidad sería del 29 por ciento.

A la vista de estos datos, podemos comprender con toda claridad por qué hubo de transcurrir dos tercios de siglo antes de que los antiguos se decidieran a construir cuadrirremes y quinquerremes en cantidad. Ciertamente pudo influir también el apego de los atenienses al trirreme, por su caracter tradicional. En efecto, las naves que vencieron en Salamina no podían caer en el olvido con tanta rapidez. El análisis no explica la causa de que los cuadrirremes y quinquerremes ganaran en popularidad a partir del 330 a. de C., y condujeran a la consiguiente carrera de armamentos navales.

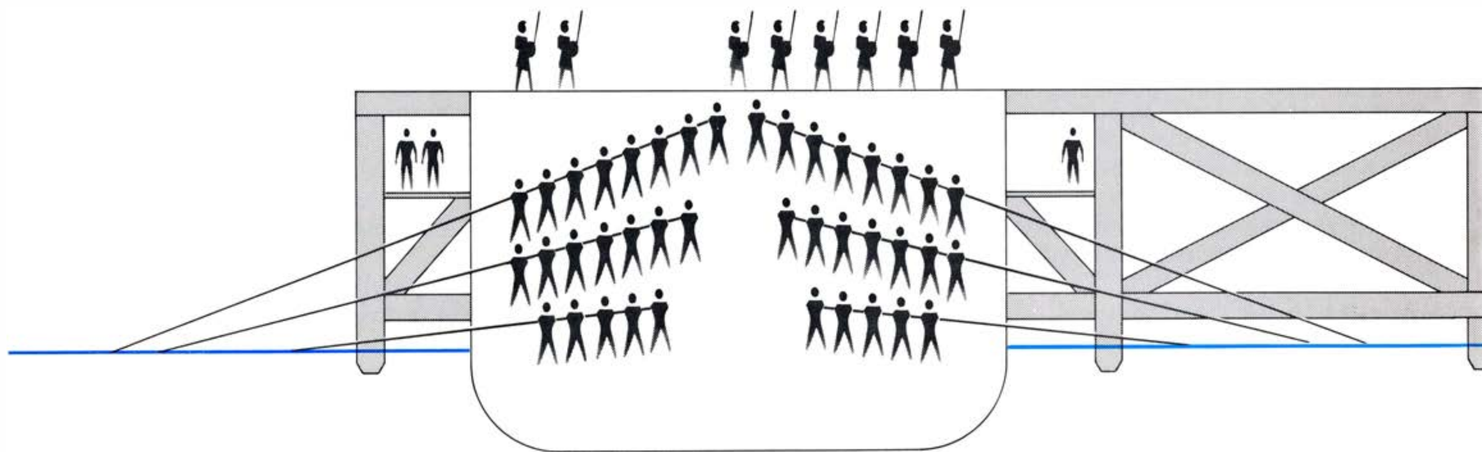
Las galeras de seis órdenes aparecieron en Macedonia y Siracusa, hacia el 340 o 350 a. de C., aunque realmente no se sabe casi nada acerca de ellas hasta después de la muerte de Alejandro Magno, el 323 a. de C., cuando los sucesores se repartieron el imperio. Hacia el 315 a. de C., Antigonos ordenó la construcción de naves de siete órdenes. Posteriormente, alrededor del 301, su hijo Demetrio el Sitiador tenía en su flota galeras de hasta trece órdenes de remeros. Este mismo, y con anterioridad a su muerte, las mandó construir de dieciséis. Pero sus rivales también las construyeron, quizá con el propósito de igualarlas o superarlas; aunque parezca extraño, hubo algunos que las hacían sólo de pocos órdenes. Sabemos que Lisímaco, por ejemplo, construyó una galera de ocho órdenes capaz de hacer frente a la de dieciséis de Demetrio, lo cual deja entrever que algo estaba ocurriendo en relación con el número de órdenes. Esta carrera condujo a la construcción de una galera de veinte órdenes y dos de treinta en la medianía del siglo, y de una gigantesca de cuarenta, botada posiblemente cuadro décadas después.

Estas galeras tan monstruosas son mucho menos conocidas que los trirremes, y acerca de ellas existe una gran diversidad de opiniones. Lionel Casson, de la Universidad de Nueva York, acaba de formular una hipótesis muy satisfactoria. De acuerdo con ella, la galera de seis órdenes se consiguió a base de que cada remo lo manejaran dos hombres, incluyendo los correspondientes a los bancos más bajos. Para ello fue necesario aumentar la manga del casco, y esto es independiente de si alguien lo había hecho o no anteriormente. Las galeras de más de seis órdenes exigían una nueva distribución, que se logró aumentando el número de remeros que manejaban cada remo hasta llegar a ocho, lo que representó un colosal incremento del número de órdenes. (El número de ocho hombres por remo fue el máximo posible, según se deduce de las experiencias realizadas al respecto en el Renacimiento, cuando se redescubrió esta forma de boga.)

Al haber más de dos hombres por remo, no podían manejarlo estando sentados. La nave llevaba bancos, pero al iniciar la estrepada los remeros debían levantarse, inclinarse hacia delante y, en el caso de que el remo fuera muy largo, subirse a una especie de peana para que la pala entrara en el agua. Para completar la estrepada debían deshacer los movimientos anteriores y terminar sentándose en el banco. De acuerdo con la experiencia europea obtenida sobre el particular a principios de la edad moderna, este sistema de boga tenía la ventaja de necesitar muy pocos remeros bien adiestrados, y en cualquier caso, bastaban que lo fueran los que iban más próximos a la crujía, es decir, los que asían el remo por el puño, en tanto que los demás se limitaban a aplicar su esfuerzo y a imitarlos. Con esta disposición, el mantenimiento del ritmo de boga perdía gran parte de su importancia, aunque exigía a la gente un mayor esfuerzo físico, al tener que manejar unos remos que llegaron a tener hasta 17,5 metros de longitud. De todos modos esto significaba, además, una considerable pérdida de cualidades por parte de la nave, mucho más de lo que el aumento de la manga en sí permite suponer.

A partir del momento en que se adoptó la norma de colocar varios remeros en cada remo, la distribución de los mismos en bancos se convierte en algo mucho más incierto que hasta entonces. Ahora un cuadrirreme podía ser un trirreme con dos remeros en el banco más alto, un birreme con dos hombres en cada banco, o una galera de un sólo banco por banda, y con los remos manejados por cuatro hombres. De la misma manera, la galera de Demetrio, con dieciséis órdenes de remos, podía tratarse de un birreme con ocho hombres por remo, o de un trirreme en el que mediante alguna combinación consiguiéramos el número dieciséis, como podrían ser seis hombres en los bancos altos e intermedios, y cuatro en los inferiores o más bajos. Según se desprende de los acontecimientos, fueron otros motivos distintos de la estabilidad los que obligaron a adoptar la solución consistente en hacer las naves de mucha manga y poco puntal.

Hay una serie de datos que demuestran que, en las galeras mayores que la de dieciséis órdenes de Demetrio, la forma de cómputo cambió completamente. Sabemos que a esta galera le hizo frente una de ocho órdenes construida por Lisímaco; ante este hecho, Casson enumera una serie de razones por las cuales dicha galera de ocho órdenes y todas las demás de número superior a dieciséis eran naves del tipo catamarán, es decir, estaban formadas por dos cascos debidamente unidos por medio de los yugos oportunos. En la galera de Lisímaco es muy probable que en cada casco los remeros estuvieran repartidos entre bancos situados a dos niveles y que cada remo lo manejaran cuatro hombres. Esta disposición satisface perfectamente todos los tipos de galeras con un número elevado de órdenes de remeros hasta llegar a cuarenta, en cuyo caso consistía en tres filas de bancos en cada banda del casco del catamarán, posiblemente con ocho, siete y cinco remeros, respectivamente, contados de arriba abajo, lo que representa un total de veinte remeros en cada grupo de tres bancos por banda de ambos cascos. En este caso, como tales naves tenían mucha manga, así como puntal, es posible que se hicieran con un número de niveles de remeros supe-



4000 REMEROS manejaban los remos de esta gigantesca nave de guerra de dos cascos, que mandó construir Ptolomeo IV, en Alejandría, hacia fines del siglo III a. de C. De acuerdo con la descripción atribuida al griego Calixeno y que recogen los escritos de Ateneo y Plutarco, el *tesaracóntoras*, o galera de 40 órdenes, tenía dos casos, y una eslora de 280 cúbitos (128 metros); en una ocasión embarcaron en ella 2850 soldados y 400 marineros. En este esquema, hecho por los autores de acuerdo con las indicaciones de Lionel Casson, de la Universidad de Nueva York, cada remo de los cuatro bancos superiores lo

rior a los que se daban en los trirremes. Las galeras de cuarenta órdenes de remos medían 128 metros de eslora y podían llevar hasta un total de cuatro mil remeros.

Los datos relacionados con estas galeras tan gigantescas indican que tenían mucha estabilidad; en cambio, andaban poco. De acuerdo con los resultados obtenidos con los quinquerremes, parece lógico suponer que la velocidad descendía a medida que aumentaba el número de bancos. De forma aproximada se desprende también que en ellas la relación entre la potencia y el desplazamiento era la sexta parte del correspondiente a los trirremes. Y aquí surge la pregunta: si el período anterior estuvo dominado por los trirremes y su espolón, ¿por qué en la época de Alejandro y sus sucesores se utilizaron unas galeras más lentas y con menos capacidad de maniobra?

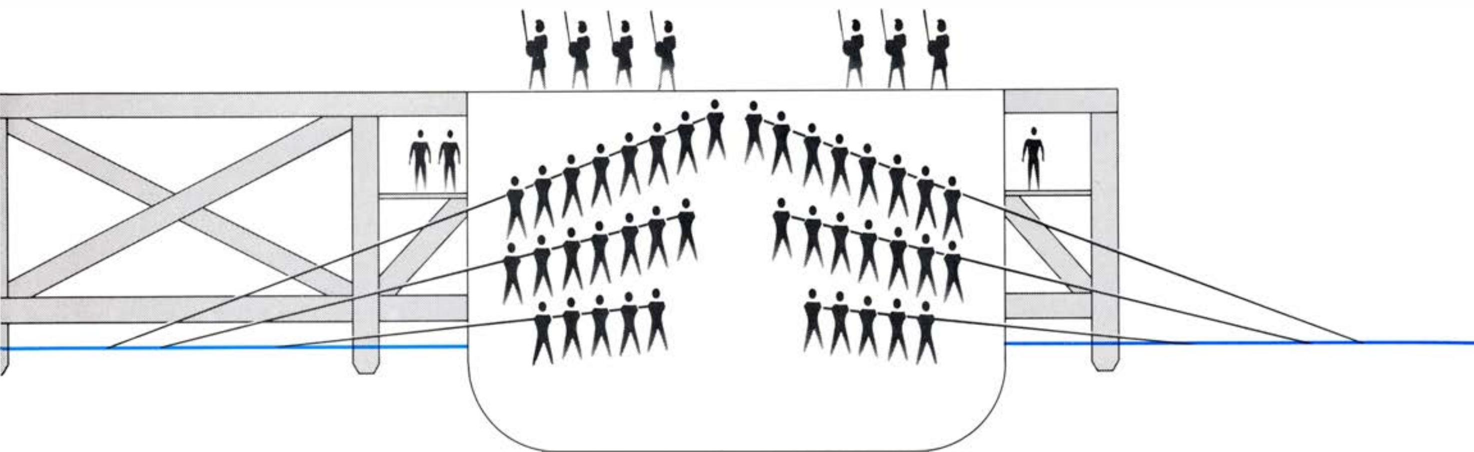
Una respuesta parcial a la anterior pregunta nos lleva a considerar la pérdida de importancia del espolón como elemento ofensivo, en beneficio del abordaje, y el empleo de ganchos o hierros que lo hicieran posible. De ahí que los quinquerremes romanos, en particular, llevaran hasta 120 soldados en cubierta. Este número significa, y lo corroboran las ilustraciones contemporáneas que han llegado hasta nuestros días, que las naves tenían mucha manga y uno o, a lo sumo, dos remeros por banco. En cambio, las galeras gigantescas con casco de catamarán llevaban un número considerablemente mayor de gente. En la galera de ocho órdenes de Lisímaco había 1200 personas, en tanto que la de 40 órdenes llevaba 2850 soldados y 400 marineros. Obviamente, los encuentros navales se decidían siempre en favor de estas supergaleras, por cuanto el número de soldados que transportaban era muy superior al de una flota de trirremes, puesto que éstos disponían sólo de 15 a 30 soldados cada uno. Además, las indicaciones espaciales reflejan perfectamente que era imposible rendir a una galera de aquella categoría con un número adecuado de trirremes, cuya potencia en conjunto fuera igual a la de ella.

Es preciso reconocer que este argumento olvida el espolón: una supergalera de ese tipo la podía hundir perfectamente el golpe de espolón de un trirreme; para evitarlo, debería ir siempre rodeada y protegida por un gran número de naves menores. Ante esto y sabiendo que la capacidad ofensiva de la misma estaba limitada por su escasa velocidad, uno se pregunta ¿por qué se construían buques de este tipo, haciéndolos cada vez de mayor tamaño? La preferencia por los buques lentos y el empleo de la táctica del abordaje nos permite deducir que hacia el año 330 a. de C. se descubrió algún sistema de neutralizar el espolón.

La solución de este verdadero rompecabezas se centra en gran parte en la utilización de la catapulta. Desde hace mucho tiempo se sabe que las galeras grandes llevaban catapultas; en su forma más moderna, es decir, más evolucionada y de mayor potencia, esas armas funcionaban por efecto de la elasticidad de los cabos hechos a base de nervios o tendones colchados, y no con piezas de madera flexible, como era usual. Las catapultas modernas aparecieron hacia el 332 a. de C. Aquel año, Alejandro montó algunas catapultas pesadas en las galeras, con el propósito de derribar las murallas de la ciudad de Tiro a la que había puesto sitio. Posteriormente, Demetrio, el mismo que aumentó los órdenes de las galeras, pasando de siete a dieciséis, las hizo instalar también a bordo de ellas.

Las catapultas montadas a bordo alcanzaban unas dimensiones realmente considerables por cuanto estaban libres de la mayoría de problemas que implicaba su traslado por tierra. Sabemos que Arquímedes hizo una catapulta para una nave que disparaba piedras de 78,5 kilogramos o dardos de 5,50 metros de largo. Estos últimos se hacían con troncos de árboles de tamaño adecuado, poniendo una punta de hierro en el extremo. En cualquier caso, e independientemente de cuál de estos dos empleara, el alcance de la máquina era de unos 200 metros, aunque usando unos proyectiles más pequeños era posible doblar dicha distancia. Se ha hablado mucho sobre el efecto producido por el impacto de una bola de piedra en el casco. Existen, empero, serias razones para dudar que fuera suficiente para atravesar el fondo de una galera y hundirla. En el peor de los casos, suponiendo que la bola cayera sobre cubierta y la atravesara, la capa de arena o grava situada en el plan y que servía de lastre amortiguaría sus efectos e impediría que atravesara el forro de la obra viva.

Sin embargo, todo el mundo se olvida de la vulnerabilidad de los soldados y de los remeros a los proyectiles de las catapultas, independientemente de que se tratara de dardos o de piedras, y en particular a partir del 500 a. de C., año en que apareció en Grecia un nuevo tipo de punta, de forma piramidal y gran poder de penetración. Según Julio César, los dardos lanzados con catapulta se clavaban hasta un pie de profundidad en la madera de roble. Obviamente esto era mucho más que suficiente, pues las galeras antiguas no tenían tanta resistencia. De acuerdo con Teofras-



manejan ocho hombres; en los bancos intermedios, van siete; y en los situados más abajo, sólo cinco. Con esta disposición, cada fila de remeros, en el sentido de la eslora, estaría formada por 100 hombres, de modo que el total de remeros sería de 4000. Según indican los textos antiguos, la longitud de los remos más largos era de 38 cúbitos (17,5 metros). Debido a su gran estabilidad, un catamarán es la nave idónea para la instalación de catapultas. De todos modos, no existe ningún indicio de que esta nave interviniera en combate, y todo parece indicar que se construyó solo para mostrarla al público.

to, las cubiertas superiores de los trirremes eran de madera de tilo, muy conocida por su poca dureza, y que se caracteriza por ser una de las maderas más ligeras que se encuentra a orillas del Mediterráneo.

Nuestro colega James F. Doyle, de la Escuela de Ingeniería Aeronáutica y Astronáutica de la Universidad de Purdue, ha efectuado ensayos con la reproducción hecha actualmente de una punta de forma piramidal y tamaño adecuado, susceptible de salir disparada con la catapulta más pequeña accionada con cabos de nervios colchados, cuya existencia se desprende de las citas contemporáneas. De ellas dedujo que la punta penetraba hasta cinco centímetros en el tilo americano, una variante de la madera antes citada, moviéndose el dardo en el momento del impacto a una velocidad igual a la mitad que poseía en el momento de lanzarlo. Esta profundidad, si tuviéramos que calificarla de algún modo, la consideraríamos como muy conservadora y la máquina que disparó el proyectil, mucho menos potente que la proyectada por Arquímedes para uso a bordo. En cualquier caso, tanto si la cubierta era de roble como si no, y las tablas tenían un grueso similar a las del forro de los costados, las galeras más grandes podían llevar un número bastante elevado de catapultas, capaces de disparar bolas de gran poder de penetración, y la mitad de la dotación normal de soldados, como mínimo.

No parece que fuera excesiva la puntería exigida a unos dispositivos de este tipo. En la Edad Media, los arqueros ingleses practicaban el tiro a una pieza de tela, extendida en el suelo, disparando desde una distancia superior a los doscientos metros. El tamaño de la tela no se conoce con exactitud, pero en cualquier caso debía ser menor que los cinco metros de la manga del trirreme. Rolfe Smith, presidente de la Asociación Nacional de Arqueros de los Estados Unidos, dice que, disparando a mano y usando el material actual es posible hacer blanco con las flechas dentro de un radio de 3 metros, a 365 de distancia. Entonces, si tenemos en cuenta que las catapultas antiguas iban montadas sobre una base regulable, probablemente los resultados fueran muy parecidos. En cualquier caso, la visión de la trayectoria de las flechas en dirección al blanco facilitaría la corrección de tiro.

Una vez atravesada la cubierta superior del trirreme, el proyectil de la catapulta tenía un amplio margen de posibilidades de acertar en algún remero, dado que éstos se encontraban muy próximos entre sí. Además, al ir los remeros en bancos situados a distintos niveles, si se empleaban proyectiles suficientemente largos y anchos, podían éstos tocar a más de un hombre a la vez; al ser disparados desde lejos, el ángulo de incidencia de los mismos era el más adecuado para ello. En el caso de que hubiera un solo hombre por remo, alcanzando a uno de ellos bastaba para desorganizar la boga de toda una banda de la galera por breves segundos, en particular si el hombre en cuestión ocupaba el banco más alto, que era el más vulnerable, y soltaba el remo de manera que cayera sobre los remos adyacentes. Sin embargo, aun en el caso de que el proyectil no hiriera a nadie, la obstrucción producida por la presencia del dardo haría perder a los remeros una estrepada, como mínimo, hasta que consiguieran arrojarlo por la borda; la demora o interrupción de la boga durante unos segundos era suficiente para producir unos efectos muy similares, por cuanto el trirreme avanzaba una longitud igual a su eslora en menos de seis segundos.

De acuerdo con todo lo expuesto, parece ser que hubo motivos suficientes para modificar el proyecto de las galeras a fin de que resultaran menos vulnerables a los proyectiles de catapulta. La sustitución de los remos manejados por un solo hombre por los de varios reducían considerablemente las posibilidades de hacer perder el ritmo de boga a toda una banda del casco, cuando se registrase una sola baja. Los seis u ocho hombres que manejaban el remo tenían fuerza suficiente para arrancar un dardo del tipo de una jabalina que hubiera hecho blanco en medio de ellos. Por otro lado, al haber menos bancos de remeros, disminuía la densidad de éstos, ofreciendo un blanco mucho menor. Esta ventaja tuvo una gran trascendencia en el proyecto de las galeras a partir de la época de Alejandro.

Por otro lado, la estabilidad contribuyó también al abandono del espolón. Normalmente, las consecuencias derivadas del golpe de espolón eran más acusadas cuando se producía en la medianía o en la aleta de la nave contraria. Por tal motivo, el mejor blanco para el espolón de una galera lo ofrecía siempre la nave que estuviera inmóvil y atravesada delante de su proa. No obstante, cuando se difundió el empleo de las catapultas la situación cambió radicalmente. Entonces la nave inmóvil estaría lanzando andanadas de proyectiles en dirección perpendicular a su eje longitudinal, y en estas condiciones, como tenía reserva de estabilidad suficiente, la puntería vendría afectada solamente por el balance de la misma, por cuanto haría variar la elevación de salida de los proyectiles, lo que a su vez

incidiría sobre su alcance. Por otro lado, la eslora de las galeras, muy grande, reduciría al mínimo los movimientos de cabeceo, que podrían introducir otros elementos de error en la puntería del disparo. Sin embargo, la nave atacante que se aproximaba, al ser diez veces más larga que ancha, ofrecía un blanco excelente a los proyectiles disparados por la nave a la defensiva.

La galera atacante debía disparar las catapultas hacia proa. Esto exigía que no cabeceara demasiado, dado que el blanco tenía tan sólo tres metros de ancho, sin incluir los remos. Para el atacante, además, el balanceo constituía un serio problema, pues al elevar las catapultas para regular el alcance levantaría más la cabeza que el extremo posterior de los proyectiles, alejando a aquella del centro de rotación de la nave.

Por efecto del balance, la cabeza de los proyectiles describía un arco hacia el costado mucho mayor que el extremo posterior de los mismos; por esa razón, de haber apuntado estando la nave adrizada, como debía ocurrir puesto que era el único momento en que podía hacerse bien, y el disparo se hacía cuando la nave tuviera una ligera inclinación, los proyectiles se desviarían hacia uno u otro lado y no darían en el blanco. En la mediania del arco que describe una nave, su movimiento de balance, o de otro tipo, es siempre mucho más rápido; no es nada fácil, pues, determinar el instante en que se encuentra perfectamente adrizada, o sea, vertical. Entonces, para una catapulta del tamaño que hemos supuesto anteriormente, el disparo hecho tan sólo 1,5 grados fuera de la vertical sería suficiente para que el proyectil errara el blanco, siempre que el trirreme atacante se encontrara a 200 metros de distancia.

De todos modos, como hemos supuesto la nave inmóvil y atravesada con respecto a la atacante, el error del que hemos hablado podría no ser muy acusado, y su influencia mínima frente a la perturbación que le crearían los movimientos de boga en sí. A pesar de la relación entre eslora y manga de la galera, en la nave atacante, durante la estrepada, la proa se levantaba un poco, para bajar seguidamente, durante la fase en que la pala del remo va por el aire. Aún con mar en calma, este leve movimiento bastaría para fallar el blanco, tratándose de un objetivo bajo y situado algo lejos. Además, si la nave atacada estuviera parada, el hecho de acertar a los remeros tendría menor importancia que de estar en movimiento.

Si hacemos un balance de cuanto hemos expuesto, llegaremos a la conclusión de que las catapultas neutralizaron en gran parte el espolón, lo que condujo a la introducción de la táctica del abordaje. De hecho, las catapultas aparecieron en el momento exacto en que se produjo el cambio de sistema de boga. Los proyectiles disparados por tales máquinas bastaban para herir a los remeros, o hundir naves, de modo que su sola presencia era suficiente para hacer perder las ganas de cualquier ataque con el espolón. Los catamaranes, al tener mucha estabilidad, eran muy adecuados para llevar catapultas. Sin embargo, queda por explicar por qué el año 250 a. de C. desaparecieron las naves de guerra de gran tamaño, usándose de nuevo los trirremes y otras galeras aún menores.

La desaparición de las galeras de gran tamaño se debió efectivamente a motivos de tipo táctico. La pesadez de éstas favoreció la intervención de naves más pequeñas y rápidas. Así, en la batalla de Chíos, el 201 a. de C., por ejemplo, unas naves muy pequeñas, conocidas como *lemboi* demostraron su



GALERA ROMANA TIPICA, probablemente un quinquerreme, en este detalle de un mosaico de principios del siglo I a. de C., fotografiado por Casson en el Palazzo Barberini, de Palestrina, cerca de Roma. Los romanos no siguieron la tradición griega, consistente en el empleo de naves veloces, sino la cartaginesa, y por ello construyeron galeras de mucha manga, capaces de embarcar a gran número de soldados. Comparados con los trirremes atenienses clásicos, que llevaban sólo 14 soldados, los quinquerremes romanos embarcaban hasta 120. En las galeras de la época final, la postiza estaba muy reforzada, de modo que se apoyaban en ella todos los remos. Motivos de índole política y táctica favorecieron la decadencia de las mayores galeras.

eficacia importunando a las naves de Rodas, mucho más pesadas. Y tras un ataque en el que lograron romper la formación de las naves grandes, los *lemboi* se pusieron entre ellas, rompiéndoles los remos o impidiendo que bogaran, con lo cual quedaban sin gobierno. En realidad, las catapultas tenían el inconveniente de que no podían disparar contra un blanco situado a menos de una distancia determinada, lo que daba lugar a la existencia de unas zonas alrededor de las galeras grandes, donde las pequeñas podían actuar con toda impunidad. Lamentablemente, no conocemos por el momento ninguna cita que nos diga que una nave pequeña se metiera entre los dos cascos de una galera grande, donde quedaría perfectamente a cubierto y así podría destrozar los remos de ambas bandas sin ningún riesgo. Desde el punto de vista táctico, la idea es excelente, pero en la práctica no era posible: la mayoría de catamaranes llevaban unos yugos de unión dispuestos entre ambos cascos, a la altura de los espolones.

Hubo también otros motivos, de índole política, que favorecieron la decadencia de las galeras de muchos órdenes. Al culminar Roma la expansión y acabar las rivalidades entre los sucesores de Alejandro, la marina quedó limitada a misiones, como la supresión de la piratería, en las que era preciso el empleo de naves pequeñas. En la última batalla naval de importancia que se disputó en la antigüedad, la de Actium, en el año 31 a. de C., las naves ligeras y rápidas de Agrippa, jefe de la flota de Octavio, derrotaron a las pesadas de Antonio. El primero utilizó algunas estratagemas que le favorecieron grandemente, tales como los ganchos de abordaje lanzados por medio de catapultas, o el empleo de proyectiles incendiarios, algunos de los cuales se disparaban con idéntica máquina. A partir de entonces, el tamaño y las necesidades de los estados exigían el empleo de unas naves muy versátiles y susceptibles de efectuar misiones de bloqueo y de vigilancia. Por otro lado, el coste excesivo de las galeras grandes, tanto en materiales como en hombres, fue otro factor en contra de ellas. Ya en aquella época incluso Roma se veía obligada a reducir sus gastos militares.

Cuando Roma entró en decadencia, la construcción naval perdió gran parte de su interés por el empleo de la energía muscular humana, de modo que, hacia el año 325, deja de hablarse ya de los trirremes. La nave típica de guerra a remo de fines del imperio romano y del período bizantino fue el dromón, que contaba como arma ofensiva principal el fuego griego, que se lanzaba a través de una especie de lanzallamas o encerrado en unos proyectiles que se disparaban con catapultas. Los dromones solían llevar los bancos dispuestos en dos niveles, y cada remo lo manejaban uno, dos o tres remeros.

Posteriormente, el mayor progreso de la galera llegó a principios de siglo XIV, cuando los italianos introdujeron el sistema de boga llamado *a zenzile,* consistente en haces de tres remos, cada uno de los cuales lo manejaba un hombre, al igual que en el trirreme primitivo. La novedad residía en que los tres remeros iban sentados en un mismo banco, dispuesto con cierta inclinación respecto al eje longitudinal de la galera, para que así no hubiera ninguna interferencia entre ellos y pudieran bogar mejor. Probablemente, el motivo principal de la introducción de este sistema fue que hacía descender el centro de gravedad de la galera; su estabilidad resultaba así muy superior a la de aquellas que llevaban los bancos situados a distintos niveles. Por aquel entonces, la aguja magnética se conocía ya en Europa, lo que permitía navegar con cierta seguridad, incluso con mal tiempo; por tal motivo, los servicios de escolta encomendados a las galeras se hacían en unas condiciones totalmente inimaginables para sus antecesores griegos.

El último cambio que experimentaron las galeras ocurrió hacia el año mil quinientos cincuenta, cuando la necesidad de armarlas con cañones de gran calibre obligó a volver a la galera de mucha manga y de varios hombres en cada remo. Y, pese a que la artillería iba instalada de manera que sólo podía disparar hacia proa, su peso obligó a introducir numerosas modificaciones, exactamente lo mismo que ocurrió dos mil años antes, cuando se montaron a bordo las catapultas. No obstante, ahora el futuro pertenecía plenamente a las naves de guerra de alto bordo, capaces de disparar andanadas por los costados y sin que los remos y los remeros representasen el menor estorbo para ello. Después de la batalla de Lepanto, ocurrida en 1571, la galera inició una rápida decadencia y tras la supervivencia excepcional en mares de poco fondo, como el Báltico, desapareció completamente. Cuando esto ocurrió, las galeras llevaban dos milenios y medio de existencia, como mínimo, en los mares del mundo occidental civilizado, lo que representa un ejemplo de la persistencia del empleo directo de la energía muscular humana en un mundo y en una cultura cada vez más dominados por las máquinas.

Proteolisis intracelular

La comprensión de los factores que regulan el recambio de proteínas aportará datos de gran interés en problemas relativos a la alimentación, el aprovechamiento de energía, el envejecimiento y las enfermedades metabólicas

Santiago Grisolía, Erwin Knecht y José Hernández-Yago

Las proteínas son macromoléculas que desempeñan funciones muy importantes en casi todos los procesos biológicos. Constituidas esencialmente por carbono, hidrógeno, oxígeno y nitrógeno, se forman a partir de unidades de bajo peso molecular (aproximadamente 75-200 dalton), los aminoácidos. Estos son ácidos orgánicos en los que un hidrógeno del átomo de carbono más próximo (carbono alfa) al radical carboxilo (− COOH) se halla sustituido por un radical amino (− NH_2). Alrededor de 20 aminoácidos están presentes en casi todas las proteínas.

La secuencia específica de aminoácidos que componen una molécula de proteína constituye su estructura primaria. La unión de los aminoácidos adyacentes en la molécula de proteína tiene lugar a través del grupo alfa carboxílico de un aminoácido y el grupo alfa amínico del otro, con pérdida de una molécula de agua y formación de un enlace covalente amídico, que recibe el nombre de enlace peptídico. La unión de varios aminoácidos constituye un polipéptido o cadena polipeptídica. La proteína puede identificarse con una sola cadena polipeptídica (piénsese en el caso de la mioglobina, una proteína del músculo) o estar constituida por varias. En este último caso, las cadenas pueden estar unidas entre sí por enlaces no covalentes (la hemoglobina de la sangre) o por el establecimiento de puentes disulfuro (-S-S-) entre los grupos -SH del aminoácido cisteína de cadenas adyacentes (valga de ejemplo la hormona proteínica insulina).

La conformación que adopta la proteína en virtud de la relación estérica entre aminoácidos adyacentes (estructura primaria) en la cadena polipeptídica se reconoce con el nombre de estructura secundaria. Esta, a su vez, condiciona la conformación definitiva de la molécula o estructura terciaria. En algunos casos, la proteína resulta de la interacción de varias cadenas polipeptídicas, o subunidades. A este nivel adicional de organización se le conoce con el nombre de estructura cuaternaria.

Las proteínas desempeñan en los organismos funciones de muy diversa índole. Así, todos los enzimas, esto es, los catalizadores de las reacciones químicas en sistemas biológicos, son proteínas. Las proteínas intervienen, además, en una gran variedad de funciones: transporte (hemoglobina, mioglobina, transferritina), almacenamiento (ferritina), respuesta inmune (anticuerpos), contracción (actina, miosina, tropomiosina), movimiento cromosómico y de cilios y flagelos (tubulina), generación y transmisión de impulsos nerviosos (colinesterasa, colinacetilasa), diferenciación y crecimiento (represores del ADN), soporte mecánico (colágeno), etcétera. Esta diversidad de funciones puede parecer sorprendente si se considera la relativa simplicidad de su estructura química. Sin embargo, debe tenerse en cuenta que el determinante crítico de la función de una proteína es su conformación o estructura espacial, que viene condicionada, en definitiva, por su estructura primaria. Habida cuenta de la elevada variabilidad de tipos de estructura primaria que presentan las proteínas, resulta comprensible la enorme diversidad de funciones que tales macromoléculas pueden llevar a cabo. Así, aunque en la molécula proteica sólo entran unos 20 aminoácidos distintos, basta considerar el número de combinaciones posible de éstos en una cadena polipeptídica de unas pocas decenas de aminoácidos para percatarse de que son prácticamente infinitas las estructuras primarias posibles.

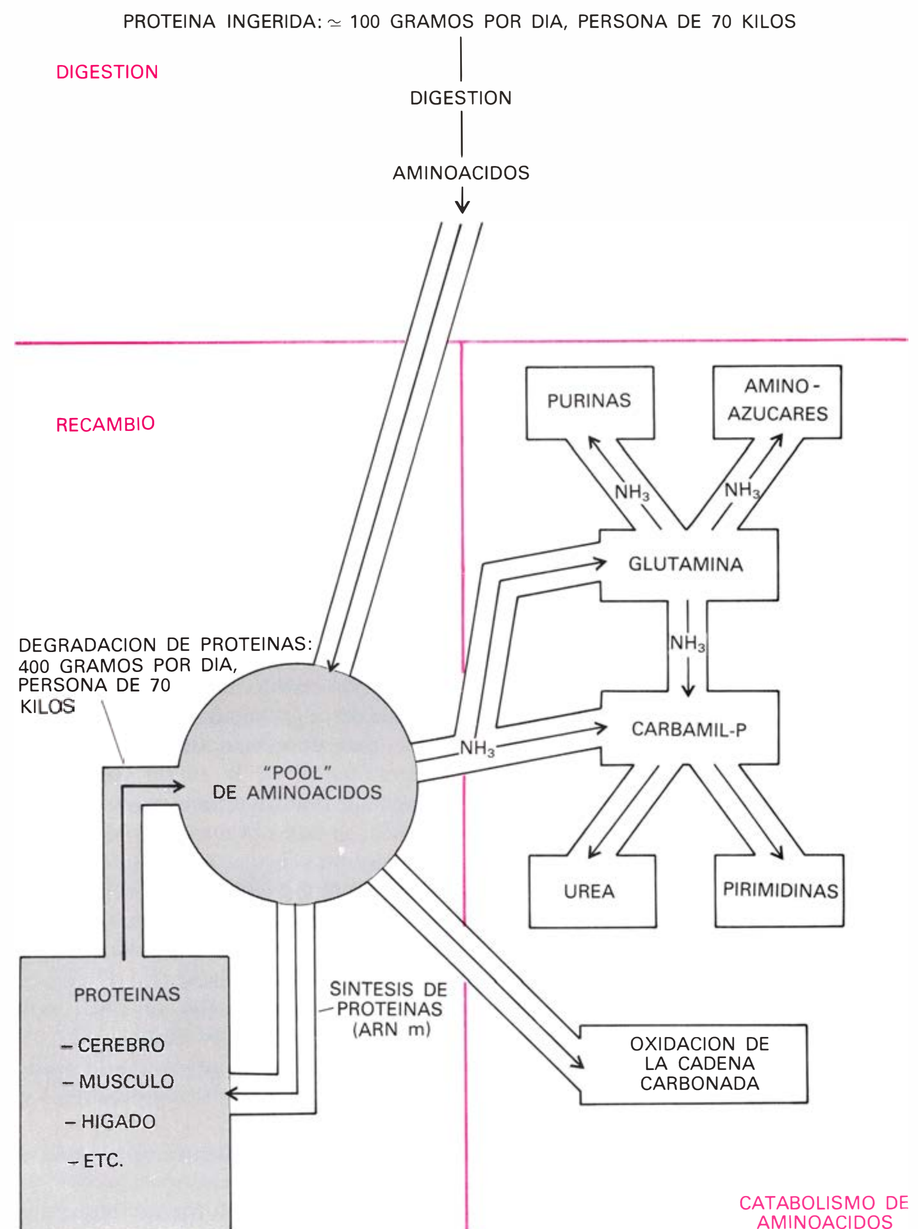
Las proteínas experimentan en las células un continuo recambio o renovación. El nivel o concentración característica que cada proteína presenta en la célula es el resultado del equilibrio dinámico existente entre las cinéticas de su síntesis y de su degradación. De hecho, en la base de los niveles nutritivos que requieren los seres vivos para su normal desarrollo, hallamos esa renovación constante a que están sometidas sus proteínas, lo que exige un suministro incesante de estos compuestos al organismo. Puede afirmarse que, si las proteínas son los alimentos más caros, se debe en gran parte a la existencia de dicho recambio. Durante más de treinta años, los investigadores han prestado especial atención a los mecanismos mediante los cuales se sintetizan las proteínas, así como a los factores celulares que regulan dicho proceso. Gracias a ese esfuerzo, poseemos hoy un conocimiento bastante preciso del mecanismo del que se sirven las células para constituir las moléculas proteicas y controlar la síntesis de proteínas específicas. En contraste franco con ello, y a pesar de que en los últimos años se ha trabajado con cierta intensidad en la elucidación de los mecanismos responsables de la degradación de esas proteínas intracelulares, lo que sabemos acerca de tales procesos desnaturalizadores es poco todavía. La elucidación de este mecanismo comprende la descripción de todos y cada uno de los pasos que conducen desde la proteína celular en estado nativo o funcional hasta sus aminoácidos constitutivos. Estos aminoácidos, producidos en la degradación, retornan al acervo ("pool") metabólico donde pueden ser reutilizados para la síntesis de nuevas proteínas, o pueden experimentar una serie de nuevas reacciones degradativas (desaminación, oxidación, etcétera) siguiendo las vías metabóicas clásicas.

Los mecanismos de la proteolisis (literalmente, degradación de proteínas) han sido objeto de extensa investigación en células eucarióticas y, en particular, en tejidos animales; menos se sabe sobre esa degradación en bacterias y en tejidos vegetales. No obstante ello,

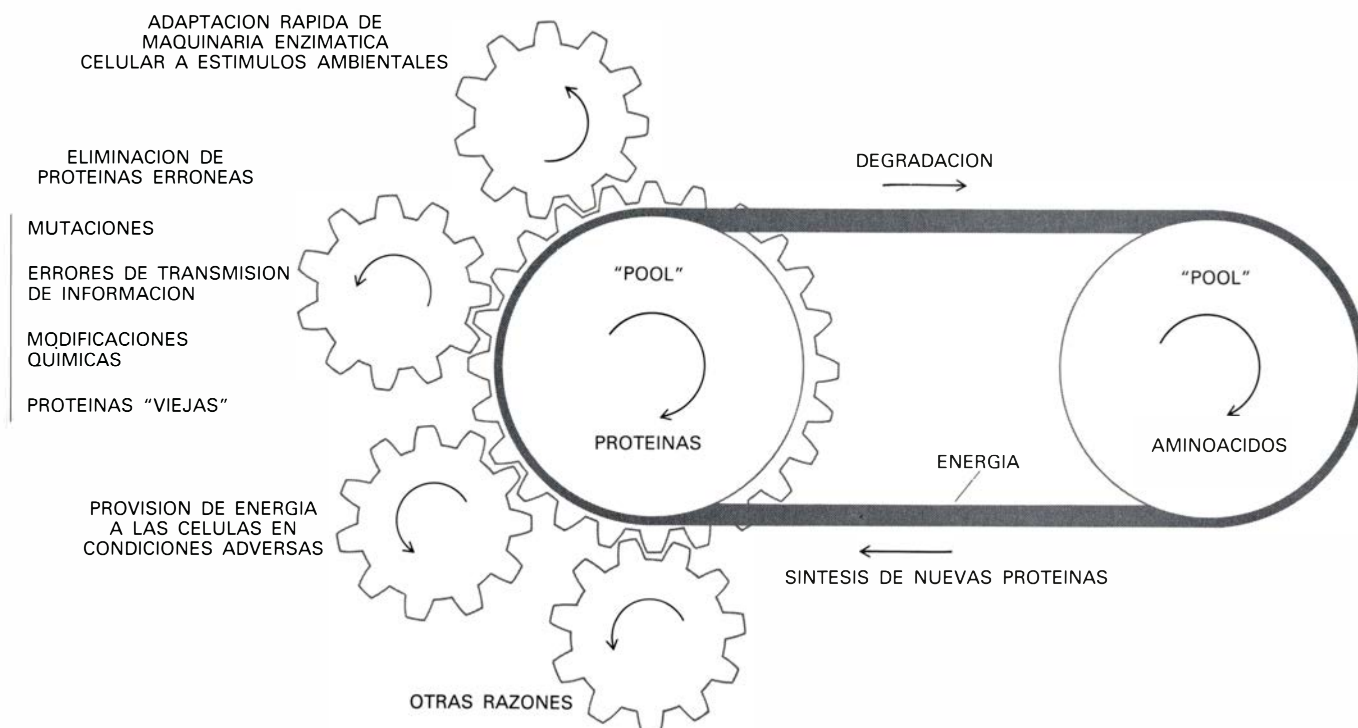
parece que podemos afirmar que los mecanismos proceden similarmente en todos los tipos celulares. El estudio de la degradación de proteínas intracelulares se relaciona, en forma más o menos directa, con otra serie de procesos biológicos fundamentales: secreción de proteínas, paso de éstas de un compartimiento celular hasta otro, recambio y muerte celular, remodelado y regresión de tejidos, degradación de proteínas exógenas (digestión del alimento, defensa contra infecciones, entre otros), proteolisis limitada a ciertas uniones peptídicas específicas en precursores mayores de otras proteínas (por ejemplo: proinsulina –precursor de la insulina–, precursores citoplásmicos de proteínas mitocondriales o con destinos diversos, etcétera). De todos estos procesos, el mejor conocido es probablemente el de la digestión de las proteínas de los alimentos. Según el modo de nutrición, los organismos se clasifican en autótrofos y heterótrofos. Los organismos autótrofos, plantas fotosintéticas y otras células, sintetizan todas sus moléculas orgánicas, directa o indirectamente, a partir de los nutrientes inorgánicos utilizando la energía que toman de la luz solar. Los restantes organismos son heterótrofos, lo que vale decir que, en última instancia, dependen para su nutrición de los organismos autótrofos al ser incapaces de sintetizar sus propias moléculas precursoras.

En este último caso, en el proceso nutritivo, se degrada el alimento a moléculas más sencillas, que puedan luego utilizarse para nuevas síntesis en las células. El lugar donde ocurre la digestión varía de un organismo a otro. Por ejemplo, entre los protozoos (amebas, paramecios, etcétera) la digestión es en su mayor parte intracelular y ocurre en unos orgánulos, los lisosomas, de los que hablaremos más adelante. En los organismos superiores, dotados de organización pluricelular, la digestión tiene lugar dentro del organismo, pero, fundamentalmente, fuera de las células (mayoritariamente en el tubo digestivo). En bacterias y hongos la digestión proteica acontece fuera del organismo, al segregar estos organismos, al medio en el que viven, enzimas digestivos. En todos estos casos es obvia la necesidad de una degradación para fabricar los productos que los organismos heterótrofos necesitan para su aprovechamiento en subsiguientes reacciones biosintéticas. Igualmente, es fácil de comprender la existencia en las células de otros procesos degradativos, tales como los implicados en procesos de defensa (digestión de bacterias invasoras) o los implicados en proteolisis limitada (por ejemplo, la degradación de las porciones N-terminales de las proteínas recién sintetizadas que precede al plegamiento tridimensional de estas proteínas). En cambio, no parece tan obvia la necesidad de que ni los organismos autótrofos ni los heterótrofos degraden las proteínas sintetizadas por sus propias células, es decir, las proteínas intracelulares. [Los organismos autótrofos sacan su nutrición del dióxido de carbono y del nitrógeno inorgánico; los heterótrofos aprovechan las moléculas orgánicas elaboradas por los autótrofos para su propio sustento.]

¿Por qué necesitan las células degradar de manera continua y extensiva sus propias proteínas? Aunque no se comprenden todavía completamente ciertos detalles del proceso de degradación de proteínas, es posible sugerir al menos tres razones que pueden justificar este proceso. La primera se refiere a la necesidad que tienen las células de eliminar proteínas alteradas que hayan perdido su carácter funcional o se hayan convertido en perjudi-



METABOLISMO DE LAS PROTEINAS en el organismo. Las proteínas que ingerimos con el alimento se desintegran, en el aparato digestivo, en sus aminoácidos constituyentes. Los aminoácidos pasan a las células ("acervo" –"pool"– de aminoácidos) y pueden ser allí utilizados para síntesis de nuevas proteínas en unos orgánulos citoplasmáticos llamados ribosomas. Cada proteína se recambia o metaboliza según una cinética precisa y específica. Los aminoácidos del "acervo" pueden experimentar, a su vez, reacciones degradativas (desaminación, etcétera), siguiendo diferentes vías metabólicas. El nivel diario de recambio de las proteínas del organismo es, en promedio, cuatro veces superior al nivel de proteína diaria ingerida. Por tanto, es obvio que existe una reutilización importante de los componentes del "acervo" de aminoácidos. Las áreas sombreadas (*abajo, a la izquierda*) indican los centros de interés de estudio de los autores.



CAUSAS QUE PUEDEN JUSTIFICAR EL RECAMBIO continuo que experimentan las proteínas en el interior celular. Aunque se desconocen los detalles del proceso, cabe sugerir al menos tres posibles razones fundadas en consideraciones teóricas: eliminación de proteínas erróneas (mutaciones, proteínas viejas, etc.), adaptación rápida de la maquinaria enzimática a estímulos ambientales y provisión de energía a la célula en condiciones adversas.

ciales para el funcionamiento normal de la célula. Existen varios procesos que pueden dar lugar a estas proteínas. Es sabido que las proteínas, una vez sintetizadas, se doblan en estructuras tridimensionales complicadas y sólo así, en su estado nativo, ejercen su función. La conformación de estas proteínas experimenta de modo continuo ligeras modificaciones, aunque con cierta frecuencia pueden llegar a alteraciones tales que la proteína tenga pocas probabilidades de recuperar su estado nativo. Se producen entonces proteínas inútiles para las células. Proteínas inoperantes pueden aparecer también por modificaciones químicas (oxidación, peroxidación, procesos de radicales libres, etcétera) y por mutaciones o errores en la transmisión de la información desde el ADN hasta el ARN y proteínas. En todos estos casos, la acumulación de proteínas inútiles o perjudiciales conduciría a una progresiva disfunción de las células que, en el mejor de los casos, determinaría un anormal crecimiento celular que dificultaría los intercambios de nutrientes y oxígeno a través de las células. Por ello constituye una ventaja para las células disponer de un mecanismo de degradación de todas esas moléculas, permitiendo que sus aminoácidos constitutivos vuelvan a utilizarse en la biosíntesis de nuevas proteínas o en otras vías metabólicas.

Una segunda función importante que puede desempeñar la degradación intracelular de proteínas es la de eliminar una parte de la maquinaria enzimática cuando ésta resulte ya innecesaria. Así, la degradación intracelular de proteínas provee a las células de una posibilidad de adaptarse a modificaciones ambientales mayor que la derivada de las modificaciones en la velocidad de síntesis o en las concentraciones de sustratos y cofactores: cuanto más rápida sea la velocidad de degradación de una proteína tanto más deprisa podrá disminuir su concentración en la célula como respuesta a estímulos ambientales. En este sentido, se han obtenido, en hígado de rata, pruebas de que muchos de los enzimas cuyas actividades controlan principalmente el flujo de sustratos a través de vías metabólicas se encuentran entre las proteínas con velocidades de degradación más rápida (es el caso, por ejemplo, de la ornitina decarboxilasa y la ARN polimerasa, que son pasos limitantes en la biosíntesis de poliaminas y del ARN.

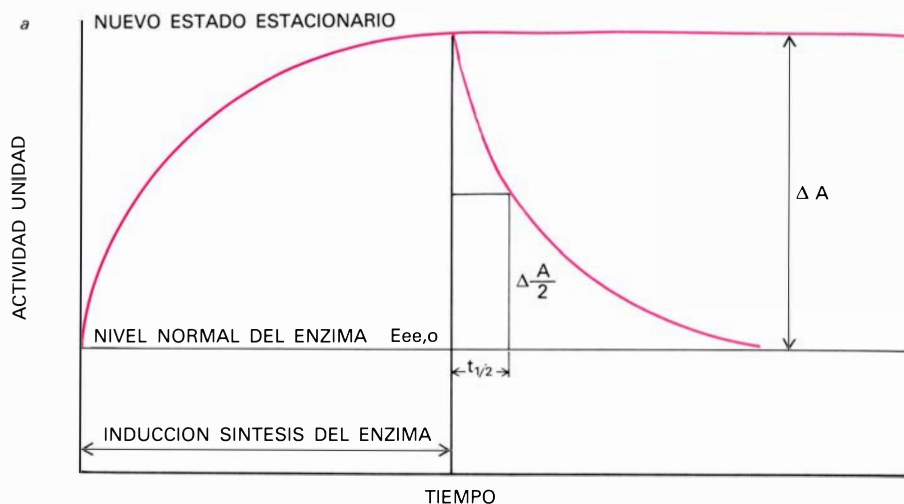
Por fin, la degradación de proteínas en condiciones adversas para las células provee a éstas de una fuente importante de energía a través de la oxidación de la cadena carbonada de los aminoácidos, así como de aminoácidos que pueden ser útiles para la síntesis de otras proteínas que sean más necesarias a las células en esas condiciones. Se sabe que, en condiciones de ayuno, aumenta la degradación de aquellas proteínas que intervienen en funciones menos esenciales, como la motilidad (por ejemplo, proteínas musculares).

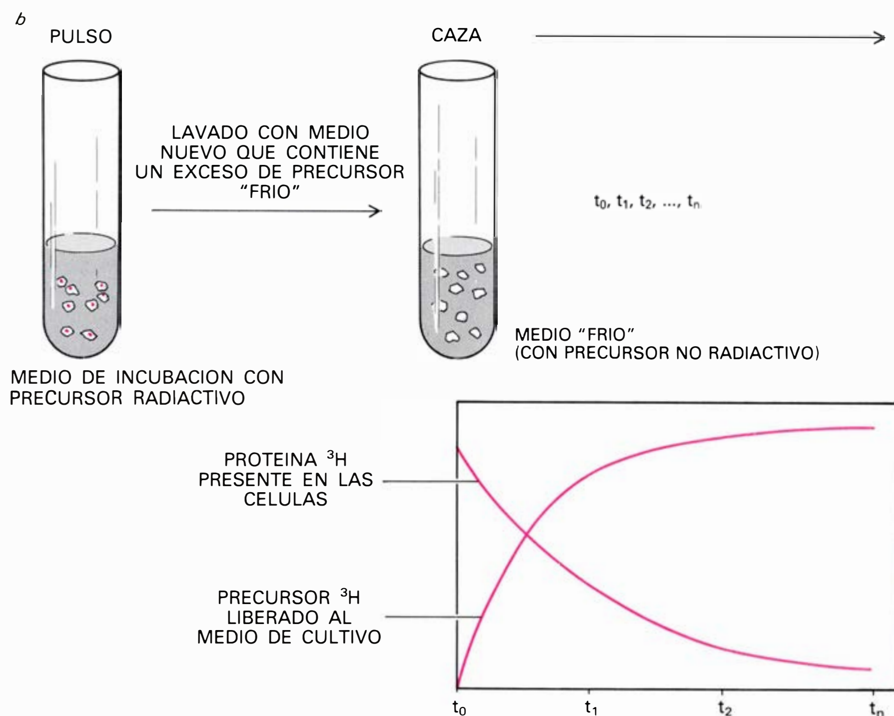
Los primeros investigadores que se ocuparon del problema del metabolismo proteico consideraron a las proteínas intracelulares, en contraposición con las proteínas exógenas, como metabólicamente inertes. La posibilidad de que existiera un recambio continuo de proteínas sólo empezó a considerarse seriamente al introducirse las técnicas con trazadores isotópicos, a finales de los años treinta. El antecedente más claro del concepto que hoy tenemos acerca del recambio de proteínas es el trabajo desarrollado, a principios de los años 40, por Schoenheimer y su grupo. En su libro póstumo (1942), Schoenheimer expresó este concepto con el término "estado dinámico de los componentes del organismo". A pesar de estos y otros trabajos, no toda la comunidad científica aceptó el hecho de la existencia de una degradación intracelular de proteínas. En la década siguiente, algunos investigadores que trabajaban fundamentalmente con organismos procariotas, consideraron que la pérdida del precursor isotópico incorporado, observada por Schoenheimer y otros, se debía a la muerte de las células o a la secreción de las proteínas sintetizadas. Sin embargo, los trabajos de Schimke principalmente, en los años 60, establecieron con toda nitidez el concepto de degradación de proteínas

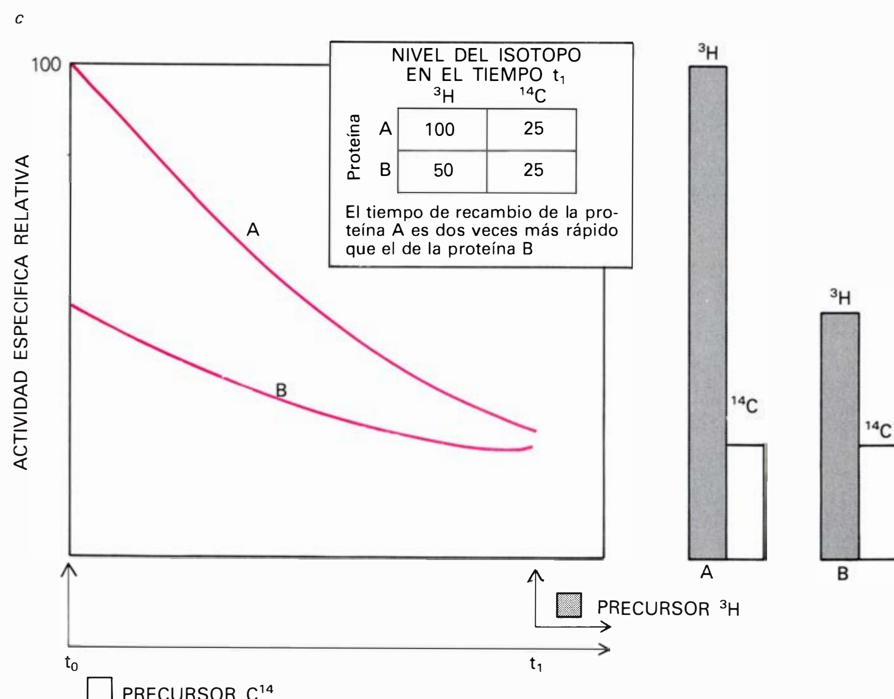
al demostrar, primero, que las proteínas de los tejidos se renovaban mucho más deprisa de lo que vivían las células y, en segundo lugar, que ciertos enzimas (por ejemplo, la arginasa de hígado de rata), para los que había podido demostrarse que existía un proceso de recambio, no se encontraban en el plasma y por tanto no cabía atribuir ese recambio a la secreción de proteínas.

A partir de entonces, se realizaron diversas mediciones de velocidades de recambio de varios enzimas (catalasa, triptófano pirrolasa, arginasa, entre otras); se estableció que había una gran variedad en las mismas, lo que conllevaba el corolario de que el proceso de degradación de proteínas debía ser específico y regulado. Los experimentos de Schimke con los enzimas arginasa y triptófano pirrolasa en hígado de rata demostraron que los cambios en la velocidad de degradación (de igual forma como ocurre con la velocidad de síntesis) son moduladores importantes en la determinación del nivel de estos enzimas. Esto despertó en los investigadores un nuevo interés por conocer el proceso de degradación de proteínas intracelulares; fruto del cual, ya en esta última década, el número de libros, artículos y congresos sobre el tema se ha incrementado muchísimo.

Obviamente, la información que pueda obtenerse acerca de los mecanismos y control en la degradación de proteínas dependerá de los métodos que se utilicen. Precisamente una de las principales razones responsables del progreso tan lento experimentado en los estudios de degradación de proteínas concierne a las dificultades metodológicas que se presentan cuando se quiere estimar cuantitativamente el







METODOS de uso corriente en la determinación de las vidas medias de proteínas celulares. Tenemos, en primer lugar, los enzimas cuyo nivel intracelular puede experimentar variaciones muy ostensibles por estímulos hormonales o fisiológicos; el descenso del nivel de enzima al retornar la célula al primitivo estado permite medir la velocidad de recambio (*arriba*). Tenemos, en el centro, el método de pulso y caza. Tras la incorporación de un aminoácido marcado (pulso), se lavan e incuban las células en un medio nuevo que contiene un exceso del aminoácido frío (caza). Se valora el recambio proteico midiendo la pérdida de radiactividad en las células o la aparición de aminoácidos radiactivos en el medio a lo largo del tiempo de caza. Abajo, el método del doble isótopo. Se administra una forma isotópica (carbono-14) de un aminácido; al cabo de 4-6 días se incorpora otra forma isotópica (tritio) del mismo aminoácido. Se sacrifica la célula o el animal a las cuatro o seis horas. Se aíslan las proteínas específicas cuyo recambio se desea valorar. El nivel de tritio asociado a proteína representará la radiactividad en el tiempo inicial; el de carbono-14, la significará en el tiempo final.

proceso. De hecho, el concepto "velocidad de degradación de proteínas" puede tener diferentes significados según sea el método empleado en su determinación. Por ejemplo, cuando se refiere a proteínas específicas, el concepto puede expresar la pérdida de proteína reconocible antigénicamente, enzimáticamente, o por cualquier otro método. En cambio, cuando se refiere a proteínas totales o grupos de proteínas, suele expresar la cinética de liberación de aminoácidos. Estos dos conceptos se suelen utilizar indistintamente, a pesar de que sólo darán valores idénticos si la primera reacción que se produce en la secuencia degradativa de una proteína es el paso limitante, esto es, el cuello de botella del proceso. En todo caso, y asumiendo en base a los pocos datos de que se dispone actualmente que la degradación de la proteína o grupos de proteínas sigue una cinética exponencial, el proceso puede describirse convencionalmente en términos de "vida media" (que simbolizaremos por *t* 1/2), que es el tiempo necesario para que la mitad de las moléculas originariamente presentes se desnaturalicen o degraden. La *t* 1/2 se relaciona con la constante de velocidad de degradación K_d, que es la pendiente en coordenadas semilogarítmicas de la recta que expresa la degradación de la proteína en función del tiempo; la vida media es inversamente proporcional a la constante de degradación. (La relación es la siguiente: $t\ 1/2 = ln\ 2/K_d$.)

La concentración de las proteínas depende tanto de su velocidad de síntesis como de su degradación y se modifica en función del tiempo. Cúmplese la relación: $dE/dt = K_s - K_dE$, donde K_s designa la constante de velocidad de síntesis (una reacción de orden 0) y E la concentración del enzima. En condiciones de "estado estacionario", no hay variación en la concentración del enzima $dE/dt = 0$). Ocurrirá, por tanto, que las moléculas que degraden se reemplazarán por la síntesis de nuevas moléculas. Se han ensayado diversos métodos para medir la degradación de proteínas. Los más utilizados han sido los siguientes: *(1)* Inducción de altos niveles de un enzima específico mediante agentes apropiados y valoración posterior de la caída de actividad enzimática. *(2)* Marcado de una o varias proteínas con un precursor isotópico y estimación de la desaparición de la proteína marcada o de la liberación de aminoácidos marcados. *(3)* Medida de un producto no metabolizable resultante de la degradación de proteínas.

PROTEINAS	FRACCION	VIDA MEDIA
Ornitina descarboxilasa	Soluble	0,2 horas
δ-aminolevulinato sintetasa	Mitocondrial	1 hora
Tirosina aminotransferasa	Soluble	1,5 horas
Triptófano oxigenasa	Soluble	2 horas
Hidroximetil-glutaril-CoA reductasa	Microsomal	2,5 horas
Fosfoenol pirúvico carboxikinasa	Mitocondrial	5 horas
Dihidroorotasa	Soluble	12 horas
Glucosa 6-fosfato deshidrogenasa	Soluble	15 horas
Ornitina aminotransferasa	Mitocondrial	19 horas
Alanina aminotransferasa	Mitocondrial	20 horas
Glutamato deshidrogenasa	Mitocondrial	24 horas
Glucokinasa	Soluble	30 horas
Catalasa	Peroxisomal	34 horas
Acetil-CoA carboxilasa	Soluble	48 horas
Adenosín trifosfatasa	Mitocondrial	60 horas
Malato deshidrogenasa	Mitocondrial	60 horas
Citocromo c reductasa	Microsomal	70 horas
α-glicerofosfato deshidrogenasa	Mitocondrial	96 horas
Arginasa	Soluble	108 horas
Citocromo b_5	Mitocondrial	120 horas
Citocromo b_5	Microsomal	150 horas
Carbamil fosfato sintetasa	Mitocondrial	185 horas
β-glucuronidasa	Lisosomal	360 horas
Lactato deshidrogenasa (isozima 5)	Soluble	384 horas
NAD glicohidrolasa	Microsomal	432 horas
Homogeneizado (promedio)		3,5 días
Núcleo (promedio)		5 días
Mitocondrias (promedio)		4 días
Membrana externa (promedio)		4 días
Membrana interna (promedio)		4,5 días
Peroxisomas (promedio)		2 días
Lisosomas (promedio)		1 día
Microsomas lisos y rugosos (promedio)		2 días
Ribosomas (promedio)		5 días
Fracción soluble o citosólica (promedio)		4 días

VIDA MEDIA de algunas proteínas de hígado de rata (valores aproximados). La velocidad de sustitución de las diferentes proteínas presenta una notable heterogeneidad. Incluso dentro de un mismo compartimiento celular (en las mitocondrias, por ejemplo), la vida media difiere de una proteína a otra. No resulta probable pues que el mecanismo autofágico pueda dar cuenta por sí solo de la degradación proteica.

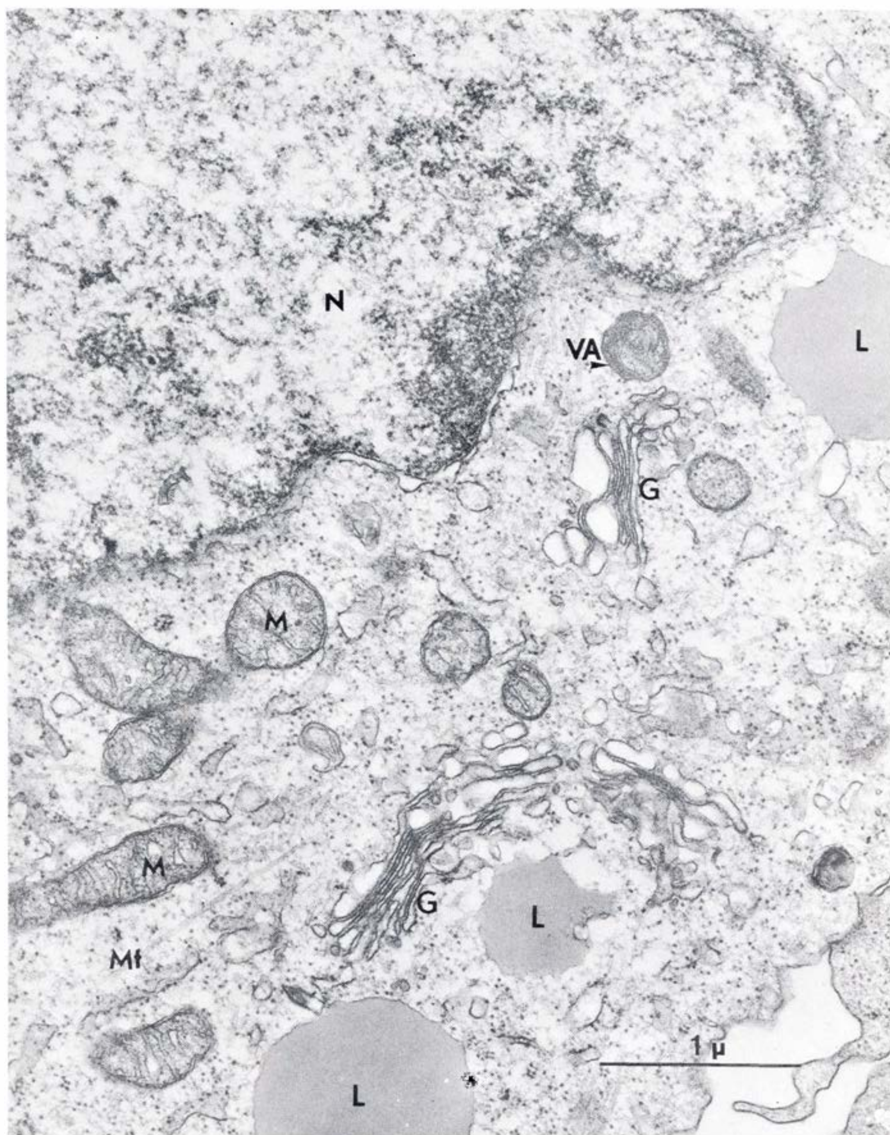
En el primero de los métodos mencionados, es decir, inducción de alta concentración de un enzima específico y valoración posterior del descenso de su actividad, el valor de K_d se deduce a partir de la pendiente de la recta obtenida al representar en coordenadas semilogarítmicas ($E_t\ E_{ee,o}$) en función del tiempo. (E_t representa la actividad enzimática a un tiempo dado, *t,* durante el período de caída de dicha actividad y $E_{ee,o}$ la actividad enzimática característica del estado estacionario en condiciones basales.) Esta técnica ha sido utilizada para medir la cinética de degradación de un cierto número de enzimas (triptófano pirrolasa en hígado de rata, inducida por administración de hidrocortisona o triptófano, o por una y otro).

En el segundo método se marcan las proteínas mediante el empleo de precursores; en éstos se ha sustituido uno o varios de sus átomos por isótopos, generalmente isótopos radiactivos: leucina (^3H), metionina (^{35}S), valina (^{14}C) y otros (por desgracia, no existen isótopos estables del nitrógeno). Se mide luego la desaparición de marca asociada a proteínas o a la aparición de marca asociada a aminoácidos procedentes de la degradación de la proteína marcada. Cuando se marcan proteínas específicas, se calcula la degradación estimando normalmente el nivel de marca asociado a la proteína purificada. Si lo que se estudia es la degradación de proteínas totales o de grupos de proteínas, se recurre con mayor frecuencia a la valoración de la marca asociada a los aminoácidos procedentes del proceso proteolítico, aunque, y ello dependerá del modelo experimental, puede medirse el nivel de marcado de las proteínas en función del tiempo. El marcado de las proteínas se lleva a cabo generalmente mediante administración del precursor radiactivo al medio de cultivo, si se trata de cultivos celulares, o por inyección vía intraperitoneal o intravenosa, cuando se experimenta con animales. Podemos marcar también las proteínas proporcionando alimento marcado radiactivamente al animal, método de raro uso porque requiere mucho tiempo y es, además, muy caro.

En todos los métodos que tienen por común denominador el marcado de una proteína con un precursor isotópico y seguimiento posterior de ella o de sus componentes, se supone que el precursor entra instantáneamente en el acervo de aminoácidos, a partir de los que se sintetiza la proteína; que el precursor es inmediatamente utilizado en la síntesis proteica y, por último, que ya

no retorna nunca al acervo, una vez administrado o inyectado. Se designa como "pulso" el tiempo de administración del precursor. Tras el pulso, la caída de radiactividad asociada a proteína o la acumulación de aminoácidos marcados constituyen una estimación de la velocidad de degradación. El punto más vulnerable de este tipo de métodos tiene que ver con la reutilización del precursor isotópico. Como se comentó antes, los aminoácidos liberados en la degradación de proteínas pueden volver a aprovecharse en nueva síntesis de proteínas. De hecho se ha calculado, tanto en células cultivadas como en tejidos, que alrededor del 50 por ciento de los aminoácidos existentes en los acervos intracelulares derivan de la degradación de proteínas. Para la célula, este proceso de reutilización supone una gran ventaja, pues de no disponer de esa facilidad habría de consumir una mayor cantidad de proteínas. Este hecho, sin embargo, presenta un problema metodológico: en tales condiciones, la medición experimental de la pérdida de radiactividad en proteínas, o de producción de aminoácidos, no sólo será función de la velocidad con que la proteína se degrade realmente, sino también de la velocidad con que el isótopo liberado en la degradación se emplee de nuevo. Se explica así que numerosas determinaciones hayan sobreestimado las *t* 1/2 de las diferentes proteínas, en especial, las que muestran un recambio más lento.

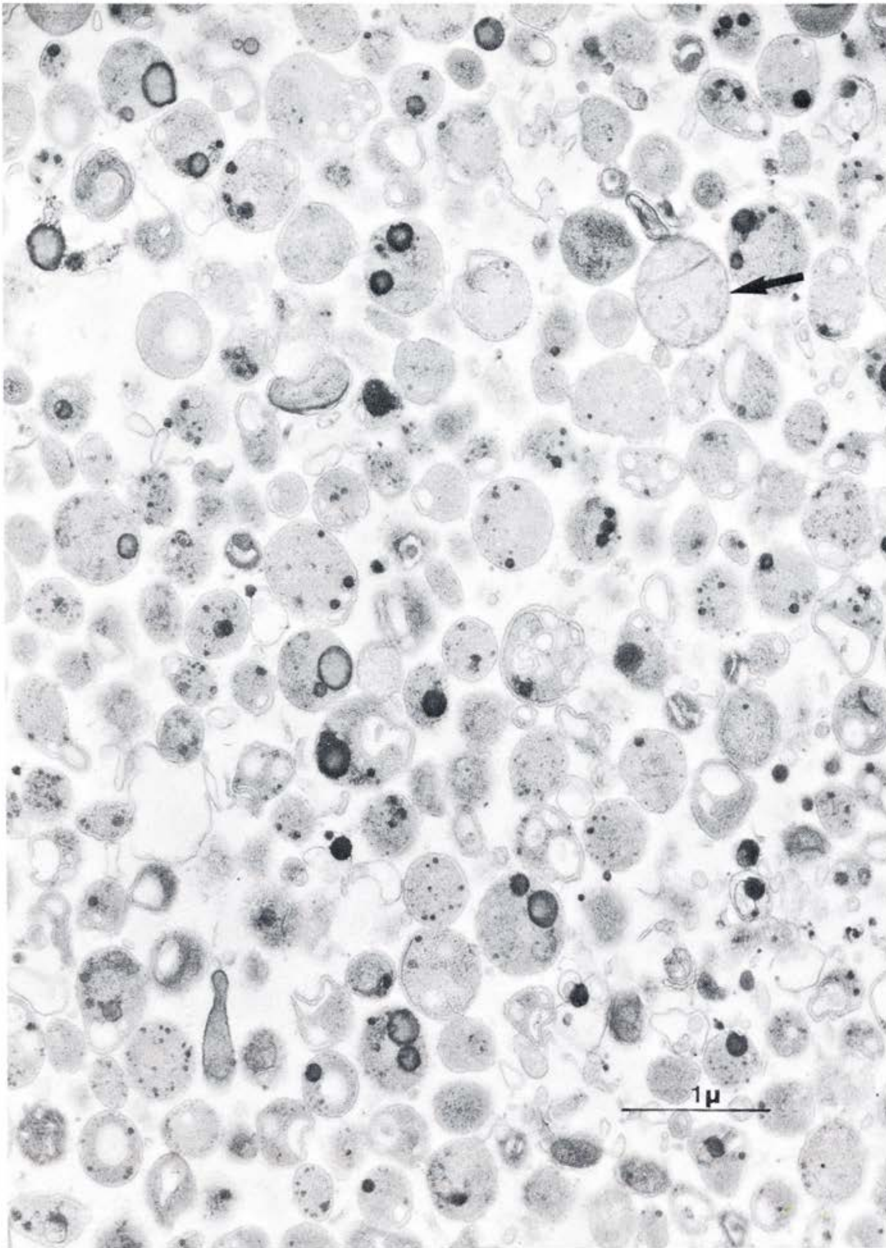
Existe una gran variedad de diseños de laboratorio que tratan de minimizar el problema de la reutilización. En los experimentos que se llevan a cabo en cultivos celulares se recurre, tras el período de pulso, a la completa eliminación del medio de cultivo, que contiene el precursor marcado, lavados extensivos y posterior incubación en otro medio de cultivo, libre del isótopo y dotado de un exceso de precursor frío (no marcado). De este modo, al crecer el acervo de aminoácidos de la célula con un exceso de precursor frío, éste compite favorablemente en su incorporación en las proteínas con el precursor marcado procedente de la degradación de proteínas sintetizadas durante el pulso. Se denomina "caza" a dicha incubación de las células en medio frío, y "experimentos de pulso y caza" a los desarrollados de acuerdo con este modelo.



LOS MECANISMOS DE LA PROTEOLISIS han sido objeto de extensa investigación en células eucariotas (provistas de núcleo diferenciado); en particular, en los tejidos animales. Menos se sabe de esa degradación en bacterias (procariotas, o carentes de núcleo) y en los tejidos vegetales. A pesar de lo cual, parece que podemos afirmar que los mecanismos proteolíticos proceden similarmente en todos los tipos celulares. La micrografía recoge la imagen de una célula cultivada. Se trata de una eucariota, dotada de núcleo (N), separado del citoplasma por una membrana; en éste se identifican las mitocondrias (M), vacuolas autofágicas o lisosomas (VA), microtúbulos (Mt), cuerpos lipídicos (L) y complejo de Golgi (G).

Hay otras vías para disminuir la reutilización de aminoácidos marcados; por ejemplo: la inhibición de la síntesis de proteínas tras el pulso y la utilización de un precursor que se degrade rápidamente en la propia célula. Cabe mencionar, en el primer caso, el empleo del antibiótico cicloheximida como inhibidor de la síntesis proteica realizada en el citosol de célula eucariótica; en el segundo, la utilización de arginina marcada con carbono-14, en el grupo guanidina, cuando se experimenta en hígado de animales ureotélicos, en los que la arginina separa dicho grupo que, a su vez, se hidroliza con la producción de urea. Pero estos dos últimos procedimientos pueden presentar nuevos inconvenientes. Por ejemplo, las concentraciones elevadas de antibióticos pueden inhibir parcialmente la degradación de proteínas. Por otro lado, los inhibidores de la síntesis proteica sólo pueden emplearse en experimentos donde el tiempo de acción del inhibidor sea breve y suficientemente bajas las concentraciones del mismo para que el efecto tóxico sea mínimo. Se sabe, además, que la inhibición de la síntesis proteica conlleva una inhibición del proceso de degradación.

En muchas ocasiones puede tener interés conocer si, en condiciones de estado estacionario, proteínas diferentes se recambian a la misma o a distinta velocidad. Para contestar a esta cuestión se ha desarrollado la técnica del "doble isótopo", que implica el recurso combinado a dos isótopos distintos de un mismo aminoácido, uno marcado con tritio, (^{3}H), y otro con ^{14}C. Se administra inicialmente al animal una for-



FRACCION LISOSOMICA de hígado de rata, obtenida a partir de un gradiente de metrizamida. Muestra la heterogeneidad morfológica de estos orgánulos autofágicos. El principal contaminante de la fracción está formado por restos de membranas. La flecha señala la presencia de una mitocondria.

ma isotópica del aminoácido (^{14}C); transcurrido cierto tiempo (de 4 a 6 días), se introduce el otro isótopo (^{3}H). El animal se sacrifica de 4 a 6 horas después de la administración del segundo isótopo y se determina la razón $^{3}H/^{14}C$ en las proteínas objeto de estudio.

Las proteínas de recambio más rápido tendrán un nivel inferior de precursor ^{14}C y un nivel superior de precursor ^{3}H que las proteínas de recambio más lento. Este método está diseñado para determinar el grado de heterogeneidad en los niveles de recambio de distintas proteínas. Si todas las proteínas se degradan con la misma velocidad, la relación $^{3}H/^{14}C$ será la misma en todas, mientras que si se degradan a distinta velocidad esta relación divergerá. Comparando esta relación con otras correspondientes a proteínas de vidas medias conocidas por otras técnicas, se puede obtener una estimación aproximada del valor de la *t* 1/2 de las proteínas estudiadas. Conviene señalar que la técnica del doble isótopo presenta serias limitaciones; citemos dos: sólo puede aplicarse a la medición de proteínas con una vida media ni demasiado corta ni demasiado larga y ciertos tejidos, como el cerebro, plantean problemas de accesibilidad por los precursores.

Reseñábamos páginas atrás un tercer método para medir la degradación de las proteínas. Por él se determina la acumulación de un producto no metabolizable procedente de la degradación de las proteínas. Se trata de una técnica de probado éxito en la medición de la degradación de las proteínas actina y miosina, en las que ciertos residuos (aminoácidos residuales) de histidina son metilados una vez incorporados en la proteína. Como la 3 metil histidina es metabólicamente inerte, y no existe en otras proteínas, la estimación del nivel de este aminoácido modificado puede revelar hasta qué punto se han hidrolizado esas proteínas. Para ello, basta con conocer el acervo total de proteínas que portan 3 metil histidina, así como la síntesis diaria de proteínas a partir de ese aminoácido. Alternativamente, se puede determinar la acumulación de la 3 metil histidina en condiciones de inhibición de la síntesis proteica. Otras proteínas en las que se producen aminoácidos especiales por modificaciones subsiguientes a la traducción son el colágeno y las histonas, con formación de hidroxiprolina en el colágeno e hidroxilisina y arginina metilada en las histonas. Para calcular la degradación de la proteína total de un tejido, podemos recurrir a este método midiendo los aminoácidos escasa o nulamente metabolizables en ciertos tejidos: valina en el hígado, fenilalanina en el corazón y tirosina o fenilalanina en el músculo esquelético.

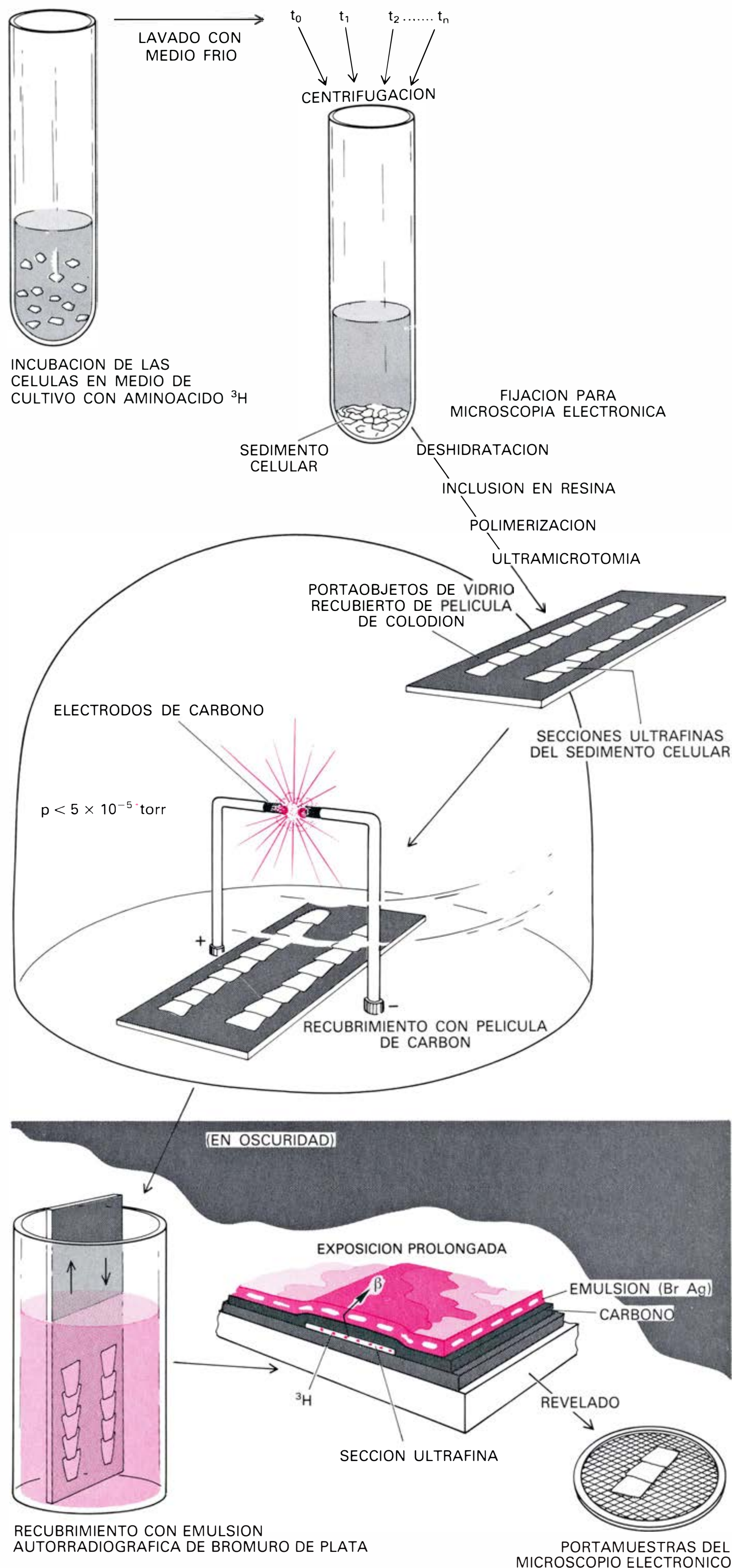
Por último, un método reciente que se presenta muy prometedor es el de incorporar en células cultivadas una proteína específica o grupos de proteínas marcadas y seguir la velocidad con que la proteína desaparece o la velocidad de liberación de aminoácidos al medio extracelular. Podemos introducir esas proteínas en el interior celular por microinyección o por fusión de las células con liposomas o con membranas vacías (fantasmas) de eritrocitos lisados y resellados que contengan la proteína marcada. (Estas últimas técnicas se están desarrollando intensivamente en nuestro laboratorio.) Las proteínas pueden ir marcadas con cualquier isótopo, pero los más apropiados son el iodo-125, ($^{125}I^{-}$), o el iodo-131, ($^{131}I^{-}$), unidos a residuos de tirosina; cuya ventaja estriba en que la tirosina iodada no se reutiliza. Este método no sólo permite obtener datos de la velocidad de degradación en determinadas proteínas, sino que, además, combinado con la autorradiografía de microscopía electrónica, podría permitir detectar puntos de degradación. Sin embargo, para que los resultados obtenidos con este método pudieran considerarse fisiológicos es indispensable que la proteína marcada que se incorpora se com-

porte de manera idéntica a la misma proteína que existe ya en la célula; es decir, que las proteínas inyectadas sean representativas del acervo total. Para ver si esto es así, una primera aproximación podría determinar si la vida media de una proteína incorporada es similar a la de la misma proteína sintetizada intracelularmente, computada por una vía diferente.

Se están ensayando otros métodos para estudiar el recambio de proteínas al objeto de soslayar los inconvenientes que comportan los métodos ya descritos. Han aparecido nuevas dificultades, por más que los métodos ganan cada día en complejidad y refinamiento. Un método diseñado en nuestro laboratorio, para valorar la posible influencia del isótopo utilizado en la estimación de la velocidad de degradación de proteínas, consiste en introducir en las células –a través de los liposomas– una mezcla de proteínas diferentes marcadas con isótopos distintos ($^{125}I^-$, ^{3}H, ^{14}C) (tantas proteínas como marcadores); se estudia luego la pérdida de cada isótopo en función del tiempo en experimentos paralelos en los que se utilizan todas las combinaciones posibles proteína-isótopo.

Se desconocen los pormenores del mecanismo de degradación de proteínas; pero se ha logrado establecer varias características generales del proceso. Algunas de estas características valen incluso para la degradación de otras macromoléculas (grasas, azúcares). Destaquemos, sumariamente, los elementos principales del proceso: *(1)* Las proteínas están sometidas a un proceso continuo de renovación. Se estima que, en un solo día, se reemplaza alrededor del 40 por ciento de toda la proteína del hígado de rata. En células cultivadas, la velocidad de substitución de sus proteínas, en condiciones adecuadas de cultivo, es muy similar. Se ha calculado que el coste calórico del recambio de proteínas en el hombre representa el 25 por ciento de las calorías necesarias en el metabolismo basal.

(2) El proceso de degradación de las proteínas tiene lugar, en su inmensa mayoría, intracelularmente. Puesto que una célula hepática viene a durar, en promedio, unos 300-400 días, no cabe atribuir a la muerte celular el 40 por ciento de degradación diaria de proteína. *(3)* Existe una notable heterogenei-



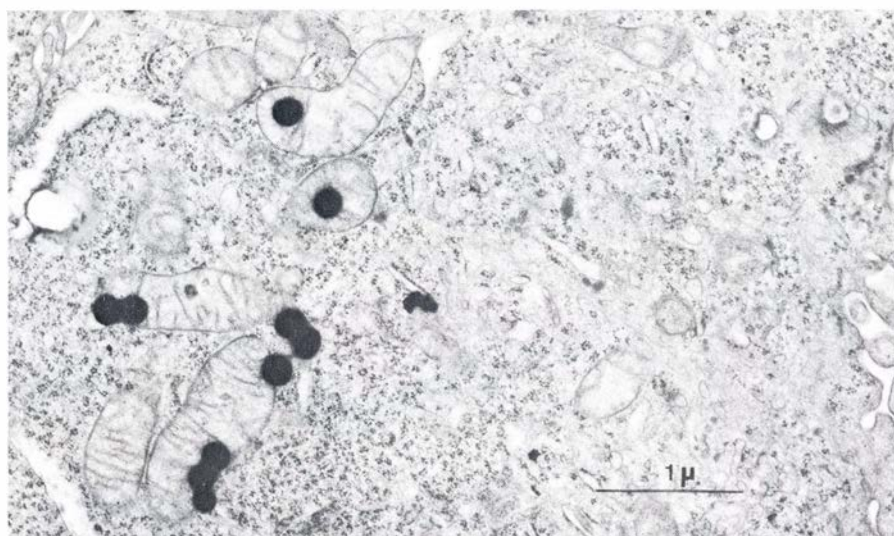
MARCADO DE PROTEINAS en células cultivadas, por administración de un aminoácido-H^3 al medio de cultivo (experimentos de "pulso" y "caza"). Aplicación posterior de la autorradiografía para revelar la presencia de proteínas marcadas en las células, mediante microscopio electrónico.

dad en la velocidad de reemplazamiento de diferentes proteínas. En hígado de rata se registran vidas medias que van desde 10 minutos para la ornitin descarboxilasa hasta 50 días para la triosa fosfato deshidrogenasa, e incluso más para otras proteínas (histonas, elastina y colágeno, que prácticamente no se recambian). Igualmente, dentro de los orgánulos subcelulares existe variabilidad para la vida media de distintas proteínas. Así en mitocondrias se encuentran valores que van desde 1 hora para la δ-amino levulinato sintetasa hasta 7,7 días para la carbamil fosfato sintetasa. Importa tener esto en cuenta a la hora de razonar el posible mecanismo implicado en el proceso de degradación de esas proteínas. Que se sepa, hay dos excepciones a lo dicho: todas las proteínas de peroxisomas y las de los microtúbulos deben probablemente recambiarse como una unidad, por tener velocidades de recambio similares todas sus proteínas.

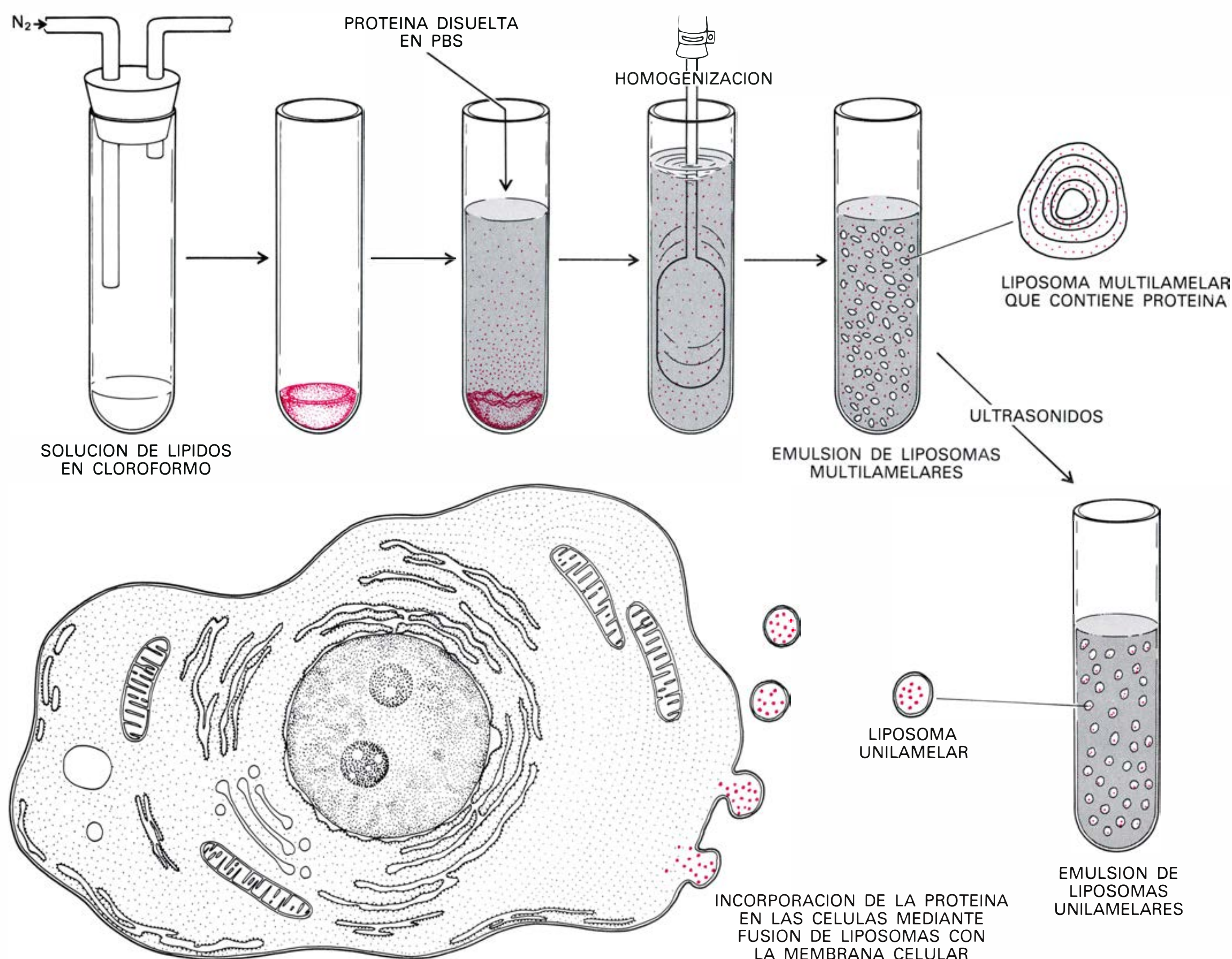
(4) La proteolisis, una vez sintetizada la proteína, parece ser un proceso al azar. Esta conclusión se basa en que la caída de radiactividad en una proteína marcada, tras la administración simple de un precursor isotópico, sigue una cinética de primer orden (esto es, una cinética que depende de la concentración existente de proteína). La explicación más probable de esta observación sería la siguiente: sintetizada la molécula proteica e incorporada en su correspondiente acervo, tiene la misma probabilidad de ser degradada que otra molécula que lleve más tiempo en el mismo. Por ello, se ha venido repitiendo de manera sistemática y rutinaria en todas las revisiones que se han realizado acerca del mecanismo de la proteolisis que se trata de un proceso al azar. Indudablemente, un mecanismo aleatorio para la degradación de proteínas satisface menos que un mecanismo más específico que reconociera las moléculas más "viejas" y, por tanto, con mayor probabilidad de alteraciones. Aunque no se dispone de prueba alguna que indique que en las moléculas que se degradan exista una acumulación mayor de modificaciones, debe señalarse, sin embargo, que la incorporación de análogos de aminoácidos (canavalina, azeditina, *p*-fluorofenil alanina, 6-fluorotriptófano) a la secuencia de aminoácidos en proteínas acelera la velocidad con la que un isótopo incorporado a la proteína se libera en forma soluble; ello hace suponer que las proteínas fuertemente modificadas tiendan a degradarse más deprisa. De ahí que nosotros consideremos que la base experimental sobre la que se sustenta la idea de una proteolisis aleatoria es muy escasa y, además, muy discutible al derivar fundamentalmente de estudios de inducción enzimática y medidas de la pérdida de actividad. Por tanto, es probable que la degradación de algunas proteínas al menos no sea al azar y quizá deban examinarse ciertas medidas antiguas con criterios más rigurosos para poner a prueba la exactitud del concepto. Aunque la verificación experimental del mismo no sea, hoy por hoy, fácil, la puesta a punto de nuevos métodos que permitan aproximaciones más precisas al proceso resultarán de enorme interés para dilucidar este punto.

(5) La velocidad de reemplazamiento de moléculas específicas varía con el estado fisiológico de la célula. Los niveles hormonales, los agentes farmacológicos y las variaciones nutricionales, en el crecimiento y en el desarrollo, pueden condicionar también la velocidad con que se degradan las proteínas. Por tanto, el ritmo de recambio de las mismas parece ser un proceso controlado de presumible interés en el crecimiento; por ejemplo, durante la regeneración del hígado, el crecimiento del mismo conlleva una reducción en la proteolisis. *(6)* En ciertos acervos de proteínas (proteínas citosólicas, por ejemplo) se ha demostrado la existencia de una correlación entre la vida media de una proteína y ciertas características de la molécula. Así, se ha señalado una relación directa entre velocidad de degradación de una proteína dada y su susceptibilidad a la inactivación por el calor y a la acción de enzimas proteasas. Se ha sugerido también una correlación entre la vida media y el peso molecular de la proteína (o el de sus subunidades), grado de hidrofobicidad (o adsorción a membranas) y carga negativa (menor punto isoeléctrico). *(7)* Se ha postulado que la proteolisis requiere energía y síntesis ininterrumpida de proteínas, por la sencilla razón de que los inhibidores metabólicos y de síntesis proteica bajan la tasa de degradación. Sin embargo, el hecho de que este tipo de inhibidores pueda actuar además a otros niveles cuestiona la existencia de una correlación de este tipo. *(8)* Se sabe, por último, que la velocidad de degradación de numerosos enzimas se modifica cuando varía la concentración de los sustratos a los que se unen y de los efectores presentes.

Hemos comentado la importancia de la degradación intracelular de proteínas, nos ocupamos después de los principios metodológicos en los que se basan los estudios acerca de este mecanismo y, por último, de ciertas características del proceso. Vamos a exponer ahora algunas ideas acerca de la posible naturaleza del mecanismo subyacente al proceso. Estamos todavía lejos de comprender en toda su extensión dicho mecanismo; por eso, en lo que sigue nos referiremos muchas veces a lo que constituyen las hipótesis de trabajo que manejamos en nuestros estudios. En particular, y puesto que es muy probable que existan diversos mecanismos para la proteolisis intracelular, hemos concentrado nuestra atención, con el fin de limitar y simplificar el problema, en la degradación de un grupo concreto, aunque muy importante, de proteínas celulares: las mitocondriales. En el proceso degradativo intervienen dos



MARCADO ESPECIFICO DE PROTEINAS sintetizadas en las mitocondrias de células eucariotas. El marcado se revela mediante una técnica de autorradiografía de alta resolución. La incorporación de leucina-tritio se llevó a cabo en presencia de cicloheximida. (La escala aparece en la parte inferior derecha de la foto.)



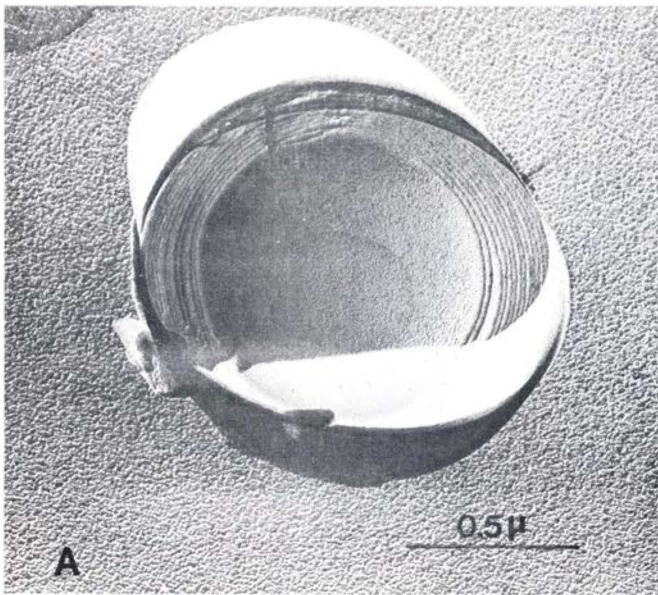
PREPARACION DE LIPOSOMAS multilamelares y unilamelares. Interacción de estos últimos con las células mediante fusión con la membrana celular y consiguiente liberación al citosol de la proteína presente en el interior de los liposomas. Otra posible vía de interacción de éstos con las células es por endocitosis (no representada), en cuyo caso los liposomas se incorporan al comportamiento lisosómico de las células. La naturaleza de los lípidos que componen los liposomas determina la vía preferente de interacción de éstos con la membrana celular. [*Véanse las fotografías de la página siguiente.*]

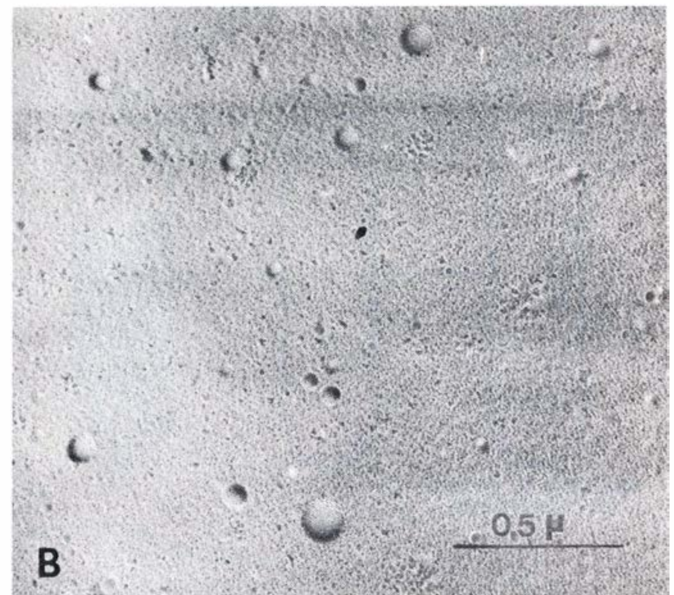
componentes: la proteína que se degrada, esto es, el sustrato de la reacción degradativa (en nuestro caso, las proteínas de la mitocondria) y el sistema degradativo que interviene en la rotura de enlaces peptídicos (es decir, los enzimas proteolíticos o proteasas). Ambos componentes pueden ser importantes a la hora de determinar la velocidad con que transcurre el proceso de recambio.

Es evidente que la actividad del sistema degradativo regulará la velocidad de degradación de las proteínas. Esta actividad dependerá de varios factores tales como la activación, o inhibición, del mismo, la traslocación de proteínas hasta el sistema degradativo y la síntesis de proteasas. La regulación en función de la actividad del sistema degradativo afectará a la degradación del conjunto de proteínas de un tejido. ¿Cuál es la naturaleza del sistema proteolítico en células animales? Es evidente que este sistema no debe ser, al menos en parte, muy específico; si cada proteína tuviese una proteasa específica que la degradase, habría de existir un enzima para desnaturalizar a otro y así hasta el infinito. Así pues, parece más lógico que exista un sistema degradativo constituido por proteasas no específicas y que las diferentes velocidades de degradación de las distintas proteínas vayan reguladas a otro nivel, como puede ser el de las propiedades específicas de cada proteína, incluyendo su estado conformacional.

¿Dónde se encuentra el sistema degradativo inespecífico en las células eucarióticas? Desde hace 25 años se sabe que la inmensa mayoría de estas células poseen unos orgánulos especializados en procesos degradativos: los lisosomas. Estos orgánulos intracelulares, limitados por una o más membranas, contienen hidrolasas ácidas (esto es, enzimas que hidrolizan múltiples moléculas y cuyo *p*H óptimo de acción es ácido para la mayoría de los enzimas). Se han encontrado en los lisosomas más de 60 hidrolasas dotadas de poder degradador de todas las macromoléculas celulares. Las proteínas no constituyen ninguna excepción. Los lisosomas poseen una gama amplia de proteasas, las "catepsinas", capaces de degradar las proteínas hasta aminoácidos y dipéptidos. Por consiguiente, esos orgánulos pueden intervenir, por lo menos, en los pasos finales del proceso de degradación, aunque es casi seguro que la proteolisis en ellos no constituye el paso limitante del proceso, ni, probablemente, el mecanismo inicial.

La membrana lisosómica aísla los enzimas hidrolíticos lisosómicos de las proteínas celulares e impide la digestión total de la célula por unos enzimas que encierran tanto peligro. Solamente algunos productos de digestión de bajo peso molecular pueden atravesar la





LOS LIPOSOMAS CONSTITUYEN una vía perfecta de introducción de diferentes sustancias en el interior celular. Se trata de vesículas lipídicas artificiales que interaccionan con las células por endocitosis o por fusión con la propia membrana que envuelve a la célula. En la micrografía electrónica de la izquierda aparece un liposoma multilamelar y, en la de la derecha, varios de ellos unilamelares. Se han estudiado mediante la técnica de criofractura.

membrana lisosómica. Los dipéptidos (o pares de aminoácidos enlazados) tienen, pues, paso franco por la membrana lisosómica. No así los tripéptidos; ni siquiera el menor de ellos: glicil-glicil-glicina. De ahí que una vez haya penetrado una proteína en el lisosoma, sólo sus productos de degradación (aminoácidos y dipéptidos) podrán salir del orgánulo atravesando la membrana. La aparente incapacidad de los lisosomas para degradar ciertos dipéptidos no plantea a las células ningún problema, porque en el citoplasma celular existen suficientes dipeptidasas para dar cuenta de la hidrólisis de cuantos dipéptidos salgan de los lisosomas procedentes de la degradación intralisosómica.

¿Cómo pueden penetrar las proteínas en el lisosoma, donde habrán de desnaturalizarse? La membrana lisosómica es impermeable al paso de macromoléculas. Sólo caben pues, dos posibilidades: o bien el lisosoma vierte sus enzimas allí donde deban degradarse las proteínas (por ejemplo, a una determinada mitocondria) o bien existe algún proceso en virtud del cual las proteínas llegan a los lisosomas. No parece que sea relevante la primera posibilidad en la proteolisis basal. Además, hay pruebas directas de la segunda, proporcionadas por la microscopía electrónica: el proceso autofágico. Este consiste en la segregación de componentes intracelulares por membranas, formándose una vacuola autofágica. Dicha estructura adquiere entonces enzimas hidrolíticos, generalmente por fusión con un lisosoma primario (llámase así el que contiene hidrolasas pero no sustratos). Se forma de este modo un lisosoma secundario, que contiene a la vez hidrolasas y sustratos y en el que ocurrirá la degradación del componente intracelular encerrado en la vacuola autofágica. Finalmente, la vacuola se transforma en "cuerpo residual", expresión que define a los lisosomas portadores de residuos indigeribles acumulados. Todo este proceso, insistimos, se basa en observaciones realizadas con el microscopio electrónico, dada la imposibilidad de reproducir de forma convincente el proceso autofágico in vitro en homogeneizados celulares.

Por lo que se refiere a las proteínas mitocondriales, la microscopía electrónica revela la presencia, en diferentes tipos celulares, de mitocondrias que atraviesan fases distintas de degradación en el interior de vacuolas. Sin embargo, conviene recordar que, puesto que en el microscopio electrónico sólo pueden observarse células fijadas y muertas, las imágenes son necesariamente estáticas, mientras que el proceso autofágico con sus diferentes fases (segregación de orgánulos, adquisición de hidrolasas, destrucción y degradación de las estructuras encerradas y formación de cuerpos residuales) constituye un proceso dinámico que se ha reconstruido a partir de esas imágenes estáticas. En este sentido, el empleo de métodos cinéticos y la utilización de métodos citoquímicos para detectar la actividad de los enzimas mitocondriales específicos y el uso de cloroquina y $ClNH_4$ para inhibir la acción lisosómica (en ambos casos posiblemente debido en parte a la elevación del *p*H intralisosómico) nos han permitido demostrar en células cultivadas la existencia de actividad enzimática de proteínas (succínico deshidrogenasa y citocromo oxidasa) en mitocondrias dentro de vacuolas autofágicas. La actividad de los enzimas desaparece progresivamente tras la supresión del tratamiento con los inhibidores lisosómicos. Esto pone de manifiesto, por un lado, que las vacuolas autofágicas intervienen en la degradación de proteínas mitocrondriales y, por otro, que las mitocondrias presentes en vacuolas autofágicas no son necesariamente mitocondrias viejas o no funcionales, sino que mantienen, al menos en parte, la actividad enzimática.

Hemos obtenido otra prueba de la participación lisosómica en la degradación de proteínas mitocondriales marcando selectivamente con leucina-^{3}H las proteínas sintetizadas en las mitocondrias en células cultivadas en presencia de cicloheximida, que, como se sabe, inhibe la síntesis proteica en el citosol pero no la que se lleva a cabo en el interior de las mitocondrias. El estudio de la distribución del precursor radiactivo en las células se llevó a cabo mediante autorradiografía de microscopía electrónica en colaboración con la doctora Martínez Ramón y se utilizaron como controles células tratadas con cicloheximida + cloramfenicol (este último es un inhibidor específico de la síntesis de proteínas en mitocondrias). Tras el pulso en presencia del precursor radiactivo y la cicloheximida, las células se incubaron en medio de cultivo "frío" sin cicloheximida; se tomaron

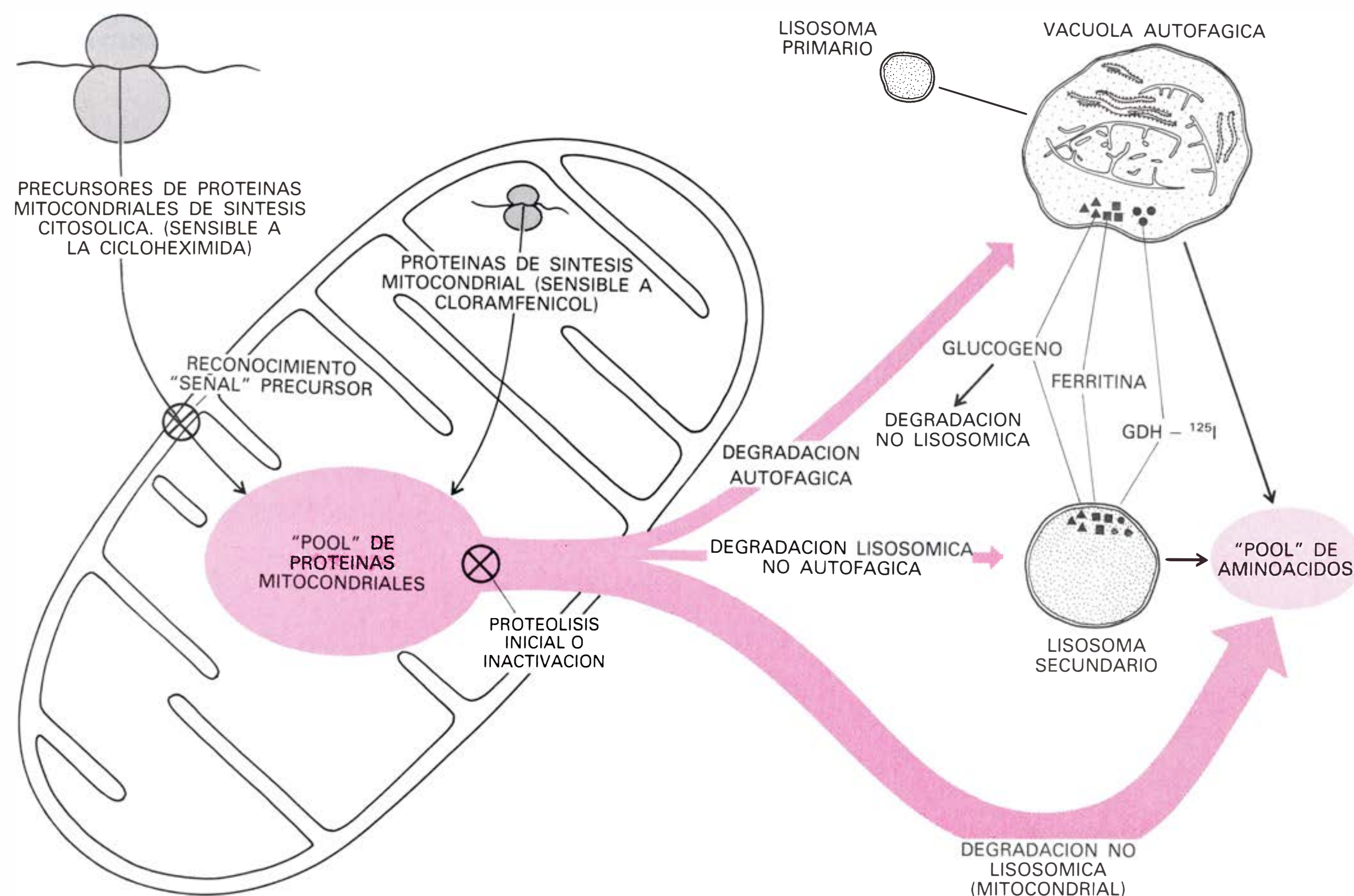
distintas muestras a diferentes tiempos del período de "caza". El análisis cuantitativo de la concentración de marcas en mitocondrias reveló la existencia de dos poblaciones principales de proteínas con vidas medias muy diferentes: 2 horas y 5 días aproximadamente. Valores similares se han obtenido con este mismo modelo experimental en varios tipos celulares, células tumorales incluidas, utilizando medidas de contador de centelleo líquido.

La enorme discrepancia entre los valores de la vida media obtenidos para las dos poblaciones de proteínas mitocondriales podría obedecer, en parte, al hecho siguiente: las proteínas de vida media más larga se corresponderían con subunidades que, en unión con otras de síntesis citosólicas, constituyeran proteínas integradas en la membrana interna mitocondrial, mientras que las de vida corta serían proteínas anormales o subunidades que, por haberse agotado el acervo de subunidades complementarias de síntesis citosólica (en ambos casos debido a la acción de la cicloheximida), no se integrarían en la membrana y se desnaturalizarían rápidamente. Cabría también que al menos una parte de las proteínas de vida media corta fueran proteínas de exportación desde la mitocondria hasta otro compartimiento celular, en contraste con las de vida media larga que corresponderían a proteínas estructurales.

Otro compartimiento celular que mostró en los experimentos autorradiográficos un marcado significativo fue el lisosómico. Ello indica que los lisosomas participan en la degradación de proteínas sintetizadas en las mitocondrias, fueran de vida media corta o larga, aunque no es posible precisar todavía en qué extensión lo hacen.

Con el fin de dilucidar hasta dónde llega la intervención de los lisosomas en esta proteolisis realizamos ensayos en células cultivadas siguiendo el mismo modelo descrito anteriormente (cicloheximida), pero utilizando cloroquina en el medio durante las primeras horas del período de caza, o bien más tarde (20 horas después del pulso). Así, al inhibir la acción de los lisosomas se acumulaban en éstos proteínas marcadas de síntesis mitocondrial, predominantemente de vida media corta, en el primer caso, y de vida media larga, en el segundo. La cuantificación de marcas en mitocondrias y en lisosomas en estos experimentos con cultivos celulares sugiere que los lisosomas están implicados en la proteolisis de un 10 por ciento aproximadamente de las proteínas de vida media corta y de un 50 por ciento de las de vida media larga. Si nos apoyamos en cálculos morfométricos desarrollados para valorar la frecuencia relativa de vacuolas autofágicas que encierran mitocondrias y de mitocondrias libres, llevados a cabo en secciones ultrafinas de células, y teniendo en cuenta los valores obtenidos para las vidas medias de mitocondrias y vacuolas autofágicas, la participación lisosómica en la degradación de proteínas de síntesis mitocondrial, de vida media larga, puede justificarse exclusivamente por el mecanismo autofágico. Este mecanismo, empero, no puede justificar la participación lisosómica en la degradación de proteínas de vida media corta; deberán existir, pues, en la célula otros mecanismos degradativos: lisosómicos no autofágicos y no lisosómicos.

¿Qué mecanismos lisosómicos no autofágicos pueden estar implicados? Para esclarecer esta cuestión hemos lleva-

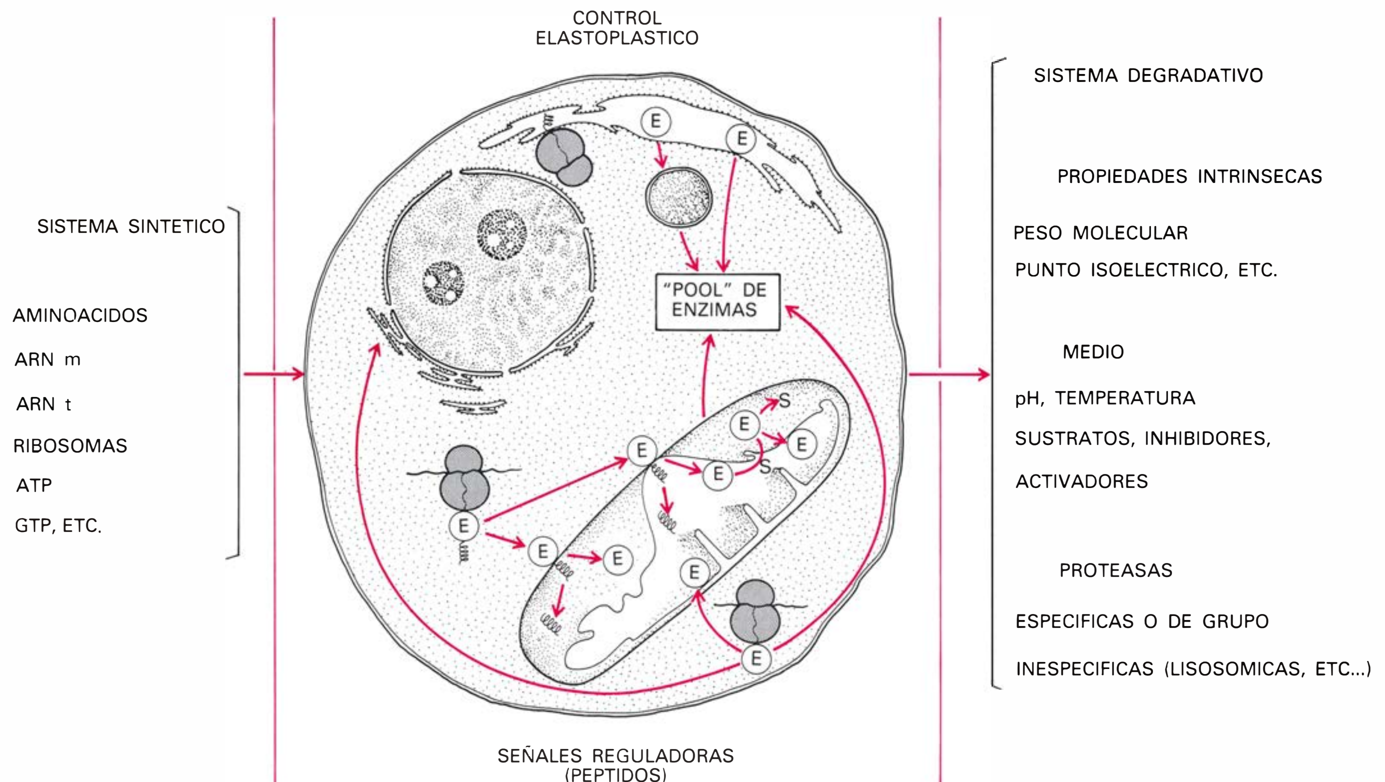


VIAS DE DEGRADACION de proteínas mitocondriales. La microscopía electrónica nos ha dado pruebas suficientes de la presencia, en diferentes tipos celulares, de mitocondrias que atraviesan fases distintas de degradación en el interior de las vacuolas. A través del empleo de métodos cinéticos y citoquímicos, los autores (que trabajan en el Instituto de Investigaciones Citológicas) han puesto de manifiesto que las vacuolas autofágicas intervienen en la degradación de proteínas mitocondriales y, además, que la mitocondrias presentes en los lisosomas mantienen, en parte al menos, la actividad enzimática.

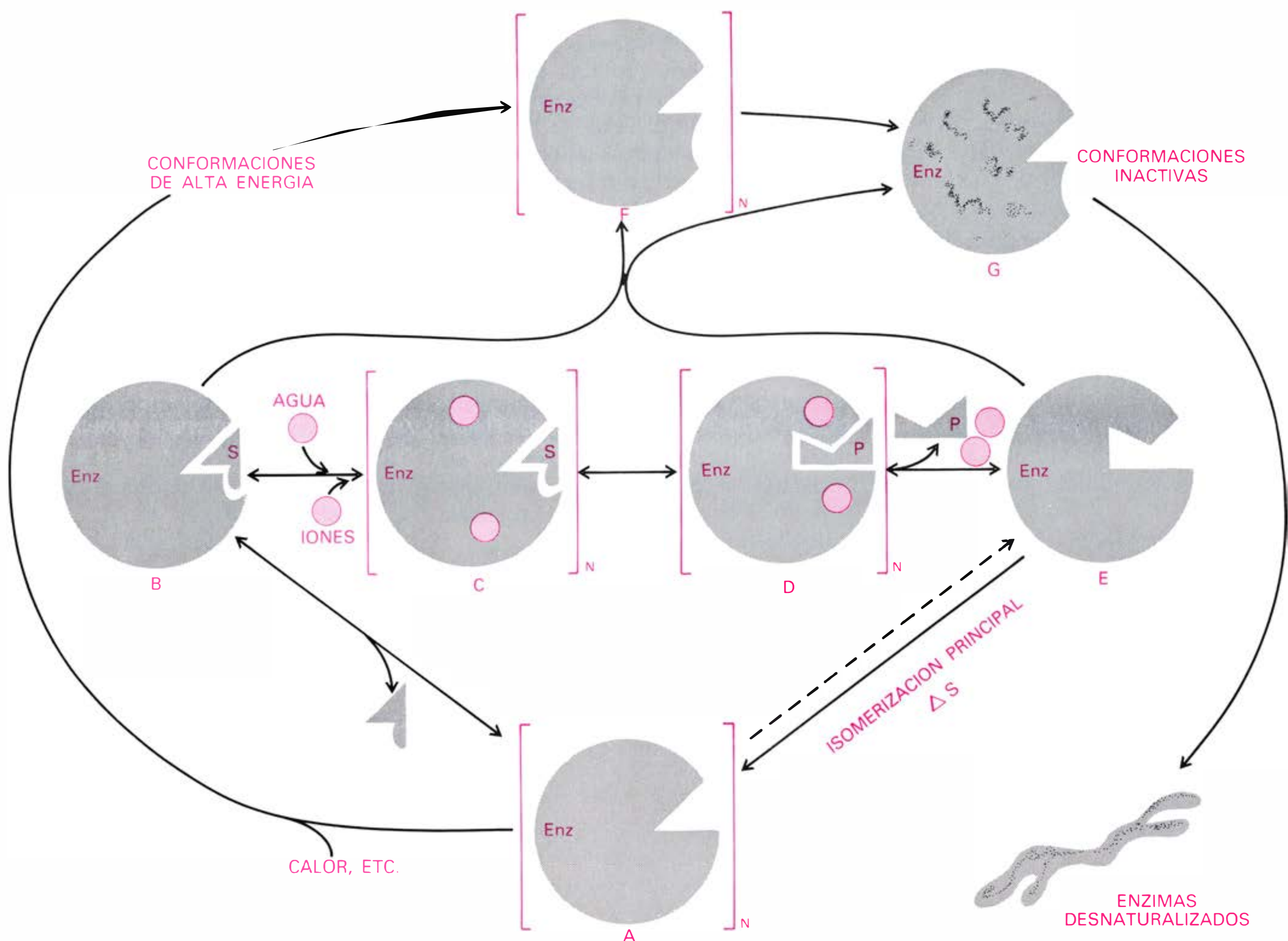
do a cabo ensayos al objeto de detectar procesos degradativos que funcionen a un nivel menos aparente que el proceso autofágico, desde el punto de vista de la microscopía electrónica. La principal dificultad estriba en que la inmensa mayoría de las moléculas individuales no pueden someterse a visualización debido a la escasa densidad de sus átomos al haz electrónico del microscopio. Como se sabe, la formación de la imagen en el microscopio electrónico requiere, entre otras condiciones, diferencias de densidad entre los distintos componentes de una estructura. Tales diferencias vienen acentuadas por la acción de determinados reactivos químicos que contienen átomos de número atómico alto (osmio, uranio, plomo), más opacos por consiguiente al paso de los electrones. El empleo de técnicas citoquímicas que permiten la identificación de moléculas individuales de glucógeno nos ha posibilitado la demostración de la existencia de una variante del proceso autofágico en la que moléculas individuales o pequeños grupos de moléculas podrían incorporarse a ciertos lisosomas: "cuerpos densos" (por la elevada densidad de su matriz) y "cuerpos multivesiculares" (lisosomas caracterizados por pequeñas vesículas en su matriz). El mecanismo de entrada podría describirse como una endocitosis intralisosómica (microautofagia).

Con el fin de verificar la validez de este proceso para proteínas, hemos incorporado ferritina (proteína detectable al microscopio electrónico en virtud del alto contenido en hierro de su grupo prostético) al citosol de células cultivadas. Se acometió esa incorporación a través de liposomas, que son unas vesículas lipídicas artificiales que se han revelado como vehículos eficaces para introducir diferentes sustancias en las células. De acuerdo con la naturaleza de sus componentes, los liposomas pueden interaccionar con las células por un proceso de endocitosis, o por fusión con la membrana celular. En el caso de endocitosis, el contenido de los liposomas pasa al compartimiento lisosómico de la célula. Si se sigue la vía fusión con la membrana celular, los liposomas transfieren su contenido al citosol de la célula. El experimento se realizó incubando cultivos celulares en una solución salina fisiológica con liposomas portadores de ferritina. Tras el período de incubación, se lavaron las células; los liposomas aún no incorporados se eliminaron, por tanto. Se incubaron luego las células en medio de cultivo. El análisis cinético llevado a cabo tras diferentes tiempos de incubación en este medio puso de manifiesto un paso de la proteína libre en el citosol a lisosomas a través de un mecanismo análogo al descrito para el glucógeno. Ensayos similares en los que el metabolito a examinar era una proteína mitocondrial, glutamato deshidrogenasa, previamente marcada con $^{125}I^-$ para permitir su localización en el microscopio electrónico mediante técnicas autorradiográficas, dieron resultados análogos. Ello nos permite sostener que este mecanismo podría considerarse igualmente válido para proteínas mitocondriales que han pasado, presumiblemente después de una proteolisis inicial en la propia mitocondria, de la mitocondria al citosol para ser degradadas.

En conclusión, los lisosomas participan en la degradación de proteínas al menos por dos mecanismos distintos: en primer lugar, por autofagia de áreas extensas de citoplasma y, en segundo lugar, por microautofagia de moléculas sencillas. En el caso de proteínas mitocondriales, el primer mecanismo sería más extensivo que el segundo. Aunque las determinaciones del grado de participación lisosómica son poco exactas (ya que no es posible precisar el porcentaje de inhibición de la degradación en lisosomas de los diferentes agentes



PRINCIPALES COMPONENTES del control elastoplástico de la síntesis y degradación de proteínas en las células. Ciertas propiedades de las proteínas parecen correlacionarse con la mayor o menor velocidad de su degradación; por tanto, resulta lógico suponer que el control de la especificidad del proceso dependa de las propiedades de la proteína como sustrato sobre el que actúa el enzima proteolítico. Junto a ese control fino, los autores proponen otro control del proceso más grosero, que determinaría los grandes cambios en la velocidad de la degradación de las proteínas y que dependería de la actividad del sistema degenerativo. Los autores se inclinan por la hipótesis de que la conformación de la proteína condicione la velocidad de proteolisis. La heterogeneidad observada en las velocidades se explicaría en el marco de un esquema general: el control elastoplástico, cuyos pormenores se ilustran luego.



HIPOTESIS ELASTOPLASTICA para explicar la velocidad de proteolisis. Supone que las moléculas de enzimas son elásticas y presentan varias conformaciones distintas (esto es, no existe un patrón proteico único "nativo"). Tras la reacción enzimática, las moléculas del enzima recuperan su configuración inicial. La conformación se modifica por cambios en el medio donde interviene. Las distintas conformaciones presentan una estabilidad diferente a los agentes, particularmente a las proteasas. Actividad y estabilidad están relacionadas. Cuando el cambio registrado en la configuración es de un grado tal que sobrepasa el límite elástico, se produce un cambio plástico. La explicación de los distintos pasos del comportamiento elastoplástico es la siguiente: A: enzimas a niveles bajos de energía. B, C y D: conformaciones de los complejos enzima-sustrato y enzima-producto. E: conformación al término de la reacción enzimática y antes de retornar a la ídem inicial. F: conformaciones de alta energía. G: conformación inactiva que conduce a la desnaturalización.

utilizados con este fin), todo parece indicar que en la célula existe, además de los mecanismos lisosómicos, otros procesos no lisosómicos más importantes, sobre todo en el caso de degradación de proteínas de vida media más corta.

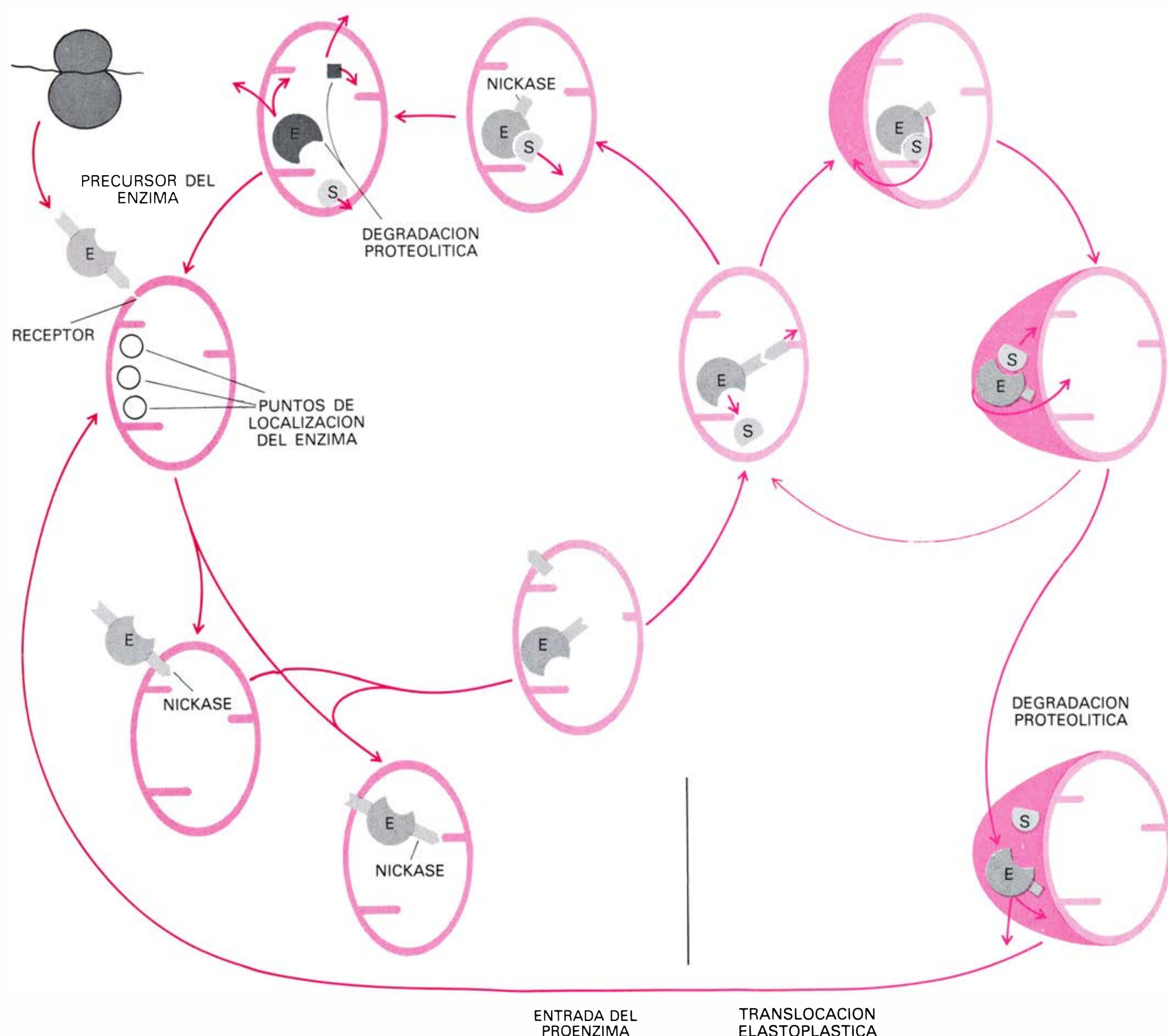
La verdad es que las proteínas mitocondriales presentan vidas medias muy dispares. Medidas de la velocidad de recambio de proteínas mitocondriales llevadas a cabo en nuestro laboratorio, y en otros, han puesto de manifiesto la existencia de una gran disparidad entre los valores correspondientes a distintas proteínas. Por tanto, a menos que se invoque la posible existencia de una heterogeneidad mitocondrial, posibilidad que estamos investigando mediante técnicas inmunocitoquímicas y morfométricas, es inevitable concluir que el mecanismo lisosómico no constituye el responsable principal de la proteolisis, habida cuenta de que la autofagia es el mecanismo lisosómico predominante y es además incapaz de explicar las diferentes vidas medias. Así pues, deben existir otros mecanismos proteolíticos en la célula.

Los lisosomas no contienen todas las proteasas celulares. Se han descrito, y es de esperar que se continúen describiendo otras, proteasas más o menos específicas en diferentes orgánulos. Por tanto, en la degradación intracelular de proteínas deben coexistir mecanismos lisosómicos y no lisosómicos para la degradación de proteínas; cada proceso de esos mostrará diferentes grados de especificidad y su importancia relativa dependerá de circunstancias distintas, regulada a través de "señales". Experimentos de proteolisis "in vitro" realizadas por nosotros con proteínas mitocondriales en hígado de rata permiten sugerir la siguiente secuencia, en términos generales, en el proceso de su degradación: pérdida de actividad enzimática, proteolisis inicial, pérdida de la capacidad antigénica y proteolisis final. A partir de esta secuencia hemos aplicado métodos inmunocitoquímicos con el fin de detectar los compartimientos celulares implicados en la proteolisis de proteínas mitocondriales específicas. No se ha detectado en lisosomas la proteína carbamil fosfato sintetasa o CPS (que supone el 20 por ciento de la proteína total de mitocondrias de hígado de animales ureotélicos), aunque sí se han aportado datos de interés en lo concerniente a la concentración de proteína antigénicamente activa presente en las mitocondrias. Los cálculos realizados para medir el efecto sobre la antigenicidad de esta proteína por parte de fijadores y otros tratamientos que conlleva la técnica inmunocitoquímica de microscopía electrónica sugieren que el número de moléculas de CPS antigénicamente activas presentes en las mitocondrias es

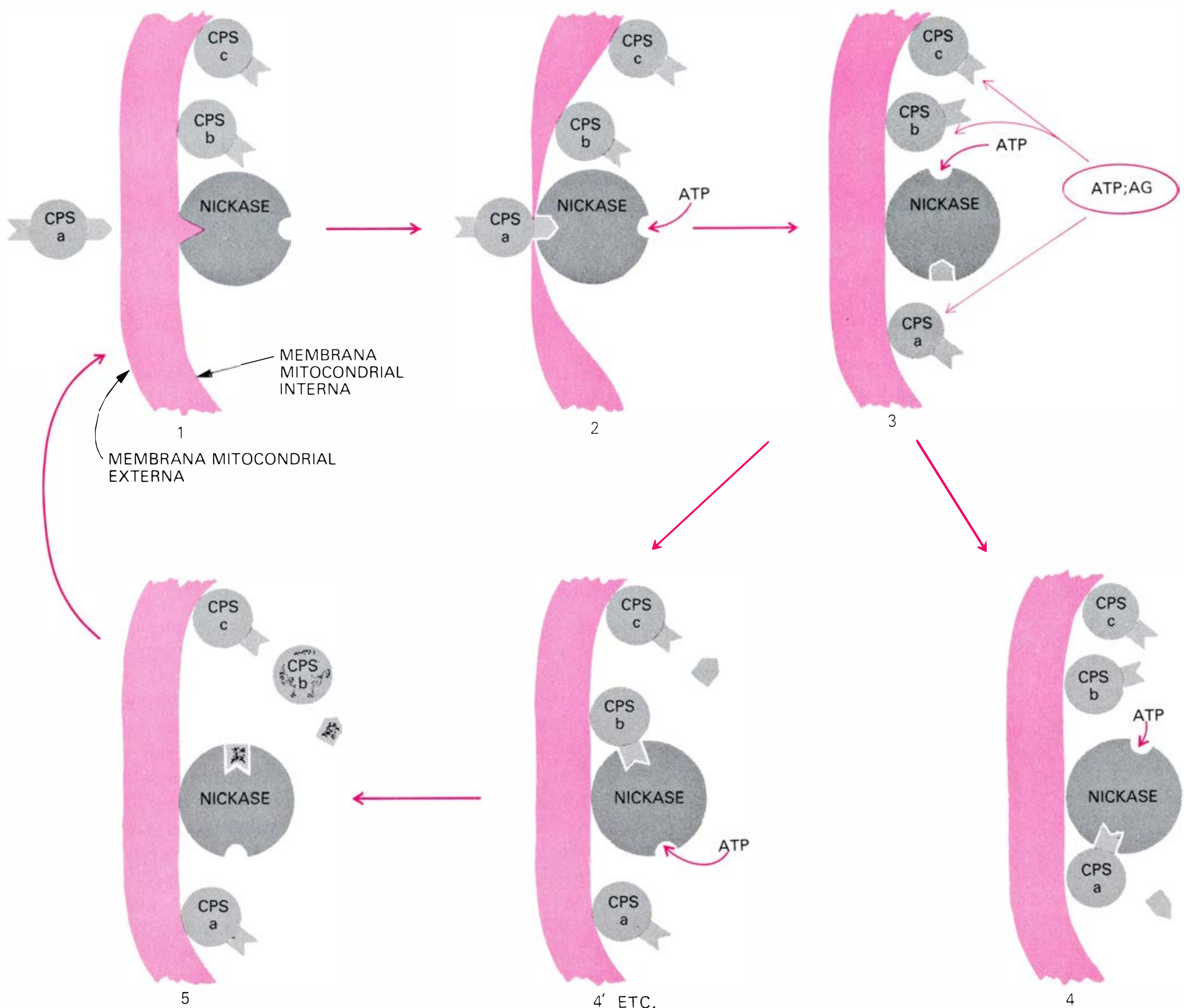
algo superior al que cabría esperar a partir de datos acerca de la actividad enzimática asociada a dichos orgánulos.

Esto nos permitió hipotetizar la posible existencia de una proteolisis inicial de la CPS en el interior de la mitocondria. Otros experimentos que completan una línea de investigación distinta iniciada por los doctores Cervera y Rubio, de nuestro laboratorio, comparando la proteolisis "in vitro" de fracciones y extractos de mitocondrias y mitoplastos (mitocondrias sin membrana externa) en presencia de extractos lisosómicos (en el caso de los extractos), a concentraciones equivalentes a la contaminación lisosómica derivada de la obtención de la correspondiente fracción y a valores de pH = 5,5 (pH óptimo de acción lisosómica) y pH neutro, evidenciaron que la proteolisis observada sólo podía justificarse asumiendo la existencia de una proteasa intramitocondrial que iniciase la degradación de las proteínas en el seno de la propia mitocondria. En este sentido, experimentos recientes llevados a cabo en colaboración con los doctores Arriaga, Soler, Timoneda y Wallace, también de nuestro laboratorio, han permitido confirmar esta hipótesis para proteínas mitocondriales específicas: carbamil fosfato sintetasa, adenosín trifosfatosa y glutámico deshidrogenasa. Para las dos primeras, la proteasa responsable de su proteolisis inicial se localiza en la membrana interna y, para la tercera, en la matriz. Actualmente se están ampliando estos estudios a otras proteínas mitocondriales.

Por otro lado, y dado que ciertas propiedades de las proteínas se correlacionan con la mayor o menor velocidad con que son degradadas, parece lógico suponer que el control de la especificidad del proceso dependa de las propiedades de la proteína como sustrato. Junto a este control fino del proceso coexistiría además el otro control, más grosero, que sería el que determinaría los grandes cambios en la velocidad de degradación de las proteínas en conjunto y que dependería de la actividad del sistema proteolítico. Es muy probable que la conformación de la proteína pueda determinar la velocidad de degradación de la misma. De acuerdo con esto, la heterogeneidad en las velocidades de degradación podría explicarse dentro de un esquema más general, la hipótesis elastoplástica, que podría justificar asimismo otros procesos biológi-



FORMACION CITOSOLICA de precursores de enzimas mitocondriales, incorporación al orgánulo, translocación elastoplástica y proteolisis intramitocondrial. Se llama citosol a la parte soluble del citoplasma que permanece después de que los orgánulos y otras partículas se han separado por centrifugación.



CICLO DE UN ENZIMA MITOCONDRIAL (en este caso, la carbamilfosfato sintetasa) de síntesis citosólica, desde su incorporación al orgánulo en forma de precursor hasta su degradación inicial en el interior de la mitocondria. AG designa, abreviadamente al acetil glutamato. *Nickase* = proteasa inicial.

cos, el del envejecimiento incluido. En forma resumida, la hipótesis supone: *(1)* Las moléculas de enzimas son elásticas y presentan varias conformaciones distintas (es decir, no existe un patrón proteico "nativo único"); *(2)* tras la reacción enzimática, las moléculas del enzima deben recuperar su conformación inicial para iniciar una nueva reacción; *(3)* la conformación del enzima se modifica por cambios en el medio en que se encuentra; *(4)* las diferentes conformaciones del enzima presentan diferente susceptibilidad (estabilidad) a toda clase de agentes (físicos, químicos y biológicos); *(5)* actividad y estabilidad parecen estar relacionadas; y *(6)* cuando el cambio conformacional resulta ser de tal grado que sobrepasa el límite elástico, se produce un cambio plástico (con menor número de grados de libertad).

Es obvio, asimismo, que el estado conformacional de la proteína pueda modificarse, entre otros motivos, por ciertas reacciones químicas que ocurren en su entorno (fosforilación, adenilación, carbamilación, acetilación, etcétera), por la disociación de la proteína en subunidades, por desnaturalización parcial, por interacción con cofactores, sustratos, proteínas, membranas (fosfolípidos). Resulta muy probable, en consecuencia, que la proteolisis venga controlada fundamentalmente a nivel de membranas (plasmática, mitocondrial, microsomal, lisosómica) a las que las proteínas con determinados cambios conformacionales –determinadas "señales"– se unirían y comenzarían las reacciones degradativas e incluso, en algunos casos, tendría lugar allí la proteolisis total. Actualmente se está abordando en nuestro laboratorio la identificación de "señales" responsables del paso de la regulación de proteínas mitocondriales de síntesis citosólica desde el citoplasma hasta el interior del orgánulo.

Así pues, estamos ante uno de los procesos básicos del metabolismo celular cuyo conocimiento, aunque todavía precario, está experimentando importantes avances. La comprensión de los factores que lo controlan aportará datos de gran interés que permitan regular la dinámica del recambio de proteínas en los seres vivos con implicaciones inmediatas en lo referente a la dilucidación de aspectos esenciales de procesos tales como el envejecimiento. Asimismo, puede contribuir a abordar otros problemas relacionados con alimentación y aprovechamiento de energía.

Juegos matemáticos

De cómo Lavinia busca alojamiento y otros problemas de muy vario carácter geométrico

Martin Gardner

Casi todos los problemas breves que este mes componen la sección son de carácter geométrico, entendiendo la geometría en sentido lo bastante amplio como para dar cabida en ella a la geometría combinatoria, la topología y la teoría de grafos. En mi próxima colaboración aportaré las soluciones a todos ellos. Las apostillas de los lectores tendrán que esperar un poco más.

1. Lavinia busca alojamiento. La recta de la ilustración de esta página representa la Avenida de la Universidad de la ciudad donde Lavinia está estudiando. Los puntos señalados con las letras *A*...*K* denotan los edificios donde residen los 11 mejores amigos y amigas de Lavinia.

Ha estado viviendo con sus padres en una ciudad cercana, pero ahora tiene la intención de mudarse a la Avenida de la Universidad. A Lavinia le gustaría encontrar una habitación situada en un lugar *L,* donde la suma de distancias a las casas de sus once amigos sea la menor posible. Suponiendo que haya alojamiento libre en el lugar óptimo, explique el lector qué debería hacer Lavinia para determinarlo, y demuestre que la suma de distancias a los once puntos señalados es verdaderamente mínima.

2. Cuerpos sólidos con simetría especular. En las figuras planas, los ejes de simetría son líneas rectas que dividen a la figura en dos mitades congruentes, cada una imagen de la otra por reflexión en el espejo. Por ejemplo, el as de corazones de la baraja tiene un eje de simetría. También tienen un eje de simetría los tréboles y las picas; los diamantes, en cambio, tienen dos. Un cuadrado posee cuatro ejes de simetría; cinco, una estrella regular de cinco puntas. El círculo tiene infinitos ejes, tantos cuantos diámetros. La esvástica y el símbolo yin-yang carecen de simetría axial.

Cuando una figura plana tiene al menos un eje de simetría, puede considerarse superponible sobre su imagen reflejada en el espejo, en el sentido siguiente. Miramos la imagen de la figura en un espejo vertical, cuyo borde inferior descanse sobre el plano horizontal donde está contenida la figura. Podemos ahora imaginar que la figura va deslizándose dentro del espejo, haciéndola girar al mismo tiempo, si es necesario, hasta llegar a coincidir con la figura reflejada. No es lícito, en cambio, darle la vuelta, permutando anverso con reverso, porque para ello sería necesario sacarla del plano horizontal y hacerla viajar por la tercera dimensión.

Un cuerpo sólido tiene un plano de simetría cuando es posible tajarlo en dos mitades congruentes, cada una imagen de la otra por reflexión especular. Las tazas de café tienen un plano de simetría, y sólo uno. La Gran Pirámide de Egipto tiene cuatro. Un cubo tiene nueve planos: tres son paralelos a otros tantos pares de caras opuestas; los seis restantes pasan por las correspondientes diagonales de estos pares de caras. Cilindros y esferas tienen infinitos planos de simetría.

Imaginemos un objeto macizo, rebanado en dos por un plano de simetría. Adosando una cualquiera de estas mitades a un espejo plano, con la cara producida al seccionarlo en íntimo contacto con la superficie de aquél, la mitad del sólido, juntamente con la imagen reflejada, recompondrán perfectamente la forma del objeto primitivo. Todo sólido dotado de un plano de simetría al menos puede ser superpuesto a su imagen reflejada, ejecutando antes, si es necesario, una rotación adecuada en el espacio.

Al analizar esta cuestión en mi libro *The Ambidextrous Universe* (Charles Scribner's Sons, 1979; hay traducción española de la edición de 1964: *Izquierda y derecha en el cosmos,* Biblioteca General, Salvat Editores S. A., 1972) afirmaba, en la página 19 (página 29 de la versión española), que cuando un cuerpo tridimensional carece por completo de planos de simetría (y así les ocurre a hélices, bandas de Möbius, nudos de margarita en bucles cerrados de soga, etcétera) será imposible superponerlo con su imagen especular sin antes imaginarlos "vueltos del revés" con un giro por la cuarta dimensión, imposible en el espacio ordinario.

¡Pero tal afirmación es falsa! Como muchos lectores del libro me hicieron notar, hay figuras sólidas que carecen totalmente de planos de simetría y que, gracias a un giro oportuno en el espacio tridimensional ordinario, pueden ser superpuestas a su imagen reflejada. Concretamente, hay una muy sencilla que podemos fabricar en un periquete, plegando una hoja cuadrada de papel. ¿Sabrá el lector construirla?

3. Una colcha de retazos deteriorada. Inicialmente, la colcha de retazos de la ilustración superior de la página siguiente, cuyas dimensiones son 9 por 12, estaba formada por 108 cuadrados de tela, de lado unidad. Algunas piezas del centro se han ajado, y ha sido preciso descoserlas. Como vemos en la figura, se han suprimido ocho cuadrados.

El problema consiste en lo siguiente: hay que descoser la colcha a lo largo de las líneas del retículo, de manera que resulten dos piezas que, cosidas entre sí, convenientemente, produzcan una colcha cuadrada de 10 por 10. Como es evidente, la colcha nueva no deberá tener agujeros. Podemos girar cada pieza a nuestra conveniencia, pero no podemos volver una de ellas del revés, porque derecho y revés de la colcha no combinan.

Aunque este problema tiene ya muchos años, es tan poco conocido y la solución tan elegante, que continuamente recibo cartas de lectores hablándome de él, ignorantes de su antiguo origen. La solución es única, y ello aunque no se exija que los cortes se ajusten a líneas del retículo.

4. Triángulos acutángulos y triángulos isósceles. Se llaman acutángulos a los triángulos cuyos ángulos interiores miden todos menos que un recto. ¿Cuál es el número mínimo de triángu-



La Avenida de la Universidad, donde Lavinia busca alojamiento

los acutángulos en que puede descomponerse un cuadrado?

Hace unos veinte años que me propuse este problema. Lo resolví entonces dando una partición del cuadrado en ocho triángulos acutángulos, como muestra la figura superior de la segunda ilustración de esta página. Al dar cuenta del problema en una columna de *Scientific American,* que puede verse reimpresa en el capítulo 3 de mis *New Mathematical Diversions* (Simon and Schuster, 1966; traducción española, *Nuevos Pasatiempos Matemáticos,* Alianza Editorial, Madrid, 1972) decía: "Durante varios días estuve convencido de que la respuesta era nueve; de pronto, se me ocurrió cómo reducirlos a ocho".

Desde aquella fecha, he recibido muchas cartas de lectores confesándose incapaces de hallar soluciones con nueve triángulos acutángulos, pero que me indicaban cómo obtenerlas con 10 o más triángulos. La figura central de la segunda ilustración de esta página deja ver cómo lograrlo con 10. Observemos que el triángulo obtusángulo *ABC* queda descompuesto en siete acutángulos gracias al pentágono de cinco triángulos. Pero si descomponemos *ABC* primero en dos triángulos, uno acutángulo y otro obtusángulo, cortándolo por *BD,* como vemos en la figura inferior, podríamos volver a servirnos del método "pentagonal" para dividir el triángulo obtusángulo *BCD* en siete acutángulos, lo que daría un total de once para el cuadrado completo. Por iteración del método, podemos lograr descomposiciones en 12, 13, 14, ..., triángulos acutángulos.

Según parece, la disección más difícil de lograr es la que emplea nueve acutángulos. Empero, es factible, como revelaré en mi próximo artículo.

Hay multitud de problemas análogos sobre división de figuras con triángulos que no se superpongan dos a dos. Aquí, mencionaré únicamente un par de ellos. La división de un cuadrado en número par arbitrario de triángulos de igual área es cosa bien sencilla pero, ¿podremos lograrlo con número impar de triángulos así? Sorprendentemente, la respuesta es negativa. Que yo sepa, el primero en demostrarlo fue Paul Monsky, en *American Mathematical Monthly* (vol. 77, n.º 2, págs. 161-164; febrero de 1970).

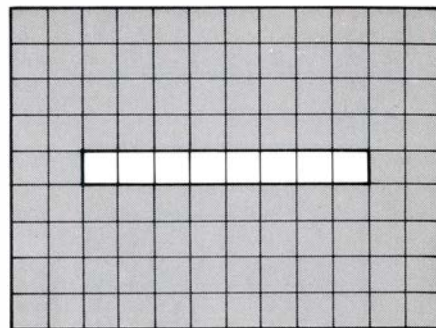
Otro curioso problema: todo triángulo es descomponible en n triángulos isósceles, con tal de que n sea mayor que 3. Puede verse una demostración, debida a Gali Salvatore, en *Crux Mathematicorum* (vol. 3, n.º 5, págs. 134 y 135; mayo de 1977). Particularmente interesante es el caso del triángulo equilátero. Es trivial descomponerlo en cuatro triángulos equiláteros (que también son isósceles), o en tres triángulos isósceles idénticos. (Hay triángulos que no pueden ser descompuestos ni en tres ni en dos triángulos isósceles, de aquí que en el teorema se requiera que n sea cuando menos 4.) ¿Sabrá el lector descomponer un triángulo equilátero en cinco triángulos isósceles? En mi próxima colaboración mostraré cómo lograrlo sin que ninguno de los cinco triángulos sea equilátero; exactamente, con uno y con dos triángulos equiláteros. No es posible que de los cinco triángulos isósceles haya más de dos equiláteros.

5. Medición con monedas de un yen. Este problema ha sido propuesto por un lector de Tokio, Mitsunobu Matsuyama, quien, además de enviarme una partida de monedas de un yen japonés, me ha explicado algunas de sus notables propiedades, no muy conocidas ni siquiera en su tierra. La moneda de un yen está hecha de aluminio puro; tiene un radio de exactamente un centímetro y pesa rigurosamente un gramo. Por consiguiente, disponiendo de un puñado de monedas de un yen y de una balanza podemos determinar el peso en gramos de pequeños objetos. También podemos servirnos de ellas para medir en centímetros distancias entre puntos del plano.

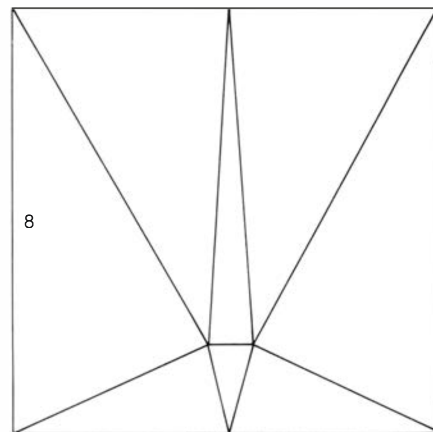
Es evidente cómo habrían de alinearse las monedas para medir distancias de número par de centímetros (dos centímetros, cuatro, seis,...) Pero, ¿servirán también para medir distancias impares (uno, tres, cinco,...)? Explique el lector cómo usar un puñado de monedas de un yen para medir distancias de un número entero cualquiera de centímetros.

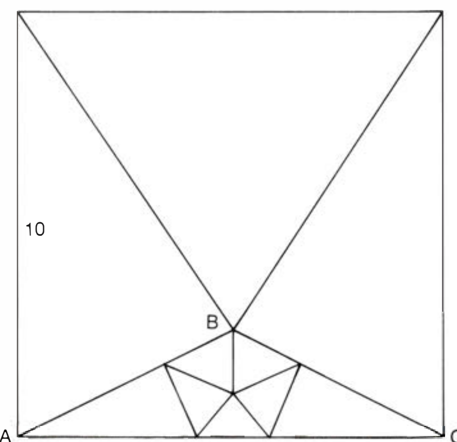
6. Un nuevo juego de coloreado de mapas. He recibido este problema directamente de su creador, Steven J. Brams, especialista en ciencias políticas, de la Universidad de Nueva York. Brams es autor de *Game Theory and Politics* (1975) y *Paradoxes in Politics* (1976), publicados ambos por la Free Press, así como de *The Presidential Election Game* (Yale University Press, 1978). Su último libro, *Biblical Games* (The MIT Press, 1980), es una sorprendente aplicación de la teoría de juegos a episodios de carácter lúdico del Viejo Testamento, donde uno de los jugadores es una deidad omnisciente.

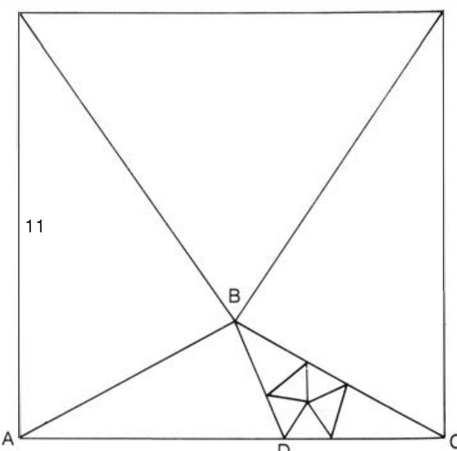
Imaginemos en el plano un mapa finito y conexo, y que disponemos de n lápices de colores distintos. El primer jugador, llamado minimizador, toma un lápiz cualquiera y colorea con él una



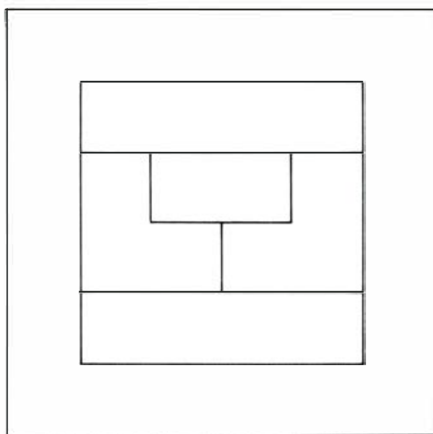
Una colcha de retazos, deteriorada







Disección de cuadrados en triángulos acutángulos



Juego de coloreado, de Brams

región del mapa, a su albedrío. El segundo jugador, maximizador, colorea entonces otra región cualquiera, usando también uno de los *n* lápices. Los jugadores prosiguen, turnándose de esta forma, iluminando región tras región con uno cualquiera de los *n* colores, aunque respetando siempre la regla de que ningún par de regiones de igual color puedan compartir tramos de frontera común. No hay inconveniente en que regiones de colores iguales se toquen en un punto.

El minimizador se propone evitar que, para dejar el mapa totalmente iluminado, hagan falta más de los *n* colores dados. El maximizador, por el contrario, se esfuerza en que sean imprescindibles al menos $n + 1$ colores. La victoria del maximizador se produce cuando alguno de los jugadores se ve imposibilitado para colorear una nueva región usando solamente los *n* lápices dados, quedando todavía regiones en blanco. Si se consigue iluminar el mapa completamente con los *n* colores, el minimizador gana la partida.

El problema, tan difícil cuan sutil, consiste en determinar el valor mínimo de *n* tal que, cualquiera que sea el mapa donde se desarrolle la partida, el minimizador pueda vencer siempre, cualquiera que sea la táctica de su oponente.

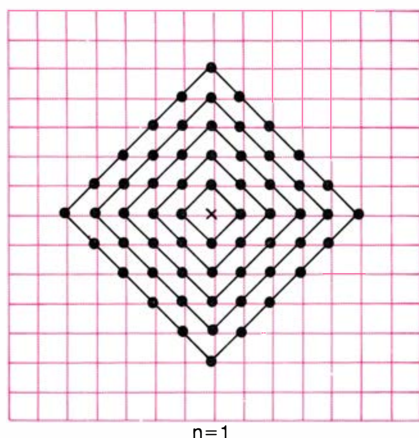
Para aclarar un poco el problema, fijémonos en el sencillo mapa de la ilustración superior de esta misma página. Vemos que en este caso *n* tiene que ser al menos 5. Desde luego, con cuatro colores podemos iluminar el mapa sin ninguna dificultad y cualquier otro mapa planar. (Así lo establece el teorema de los cuatro colores, demostrado al fin.) Pero cuando el mapa se usa como tablero para el juego de Brams, si tan sólo se dispone de cuatro colores, el maximizador puede siempre obligar a su contrario a servirse de un quinto color. Disponiendo de cinco colores el minimizador puede ganar siempre.

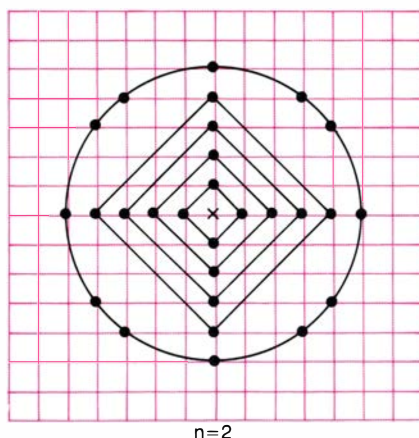
En opinión de Brams, el valor mínimo de *n* es 6. Se ha descubierto un mapa donde, con cinco colores, el maximizador tiene una estrategia que le permite vencer siempre. ¿Sabría el lector construir un mapa así, y dar la estrategia que siempre otorga la victoria al maximizador? Recuerde que el minimizador es siempre el primero en jugar, y que ningún jugador está obligado a servirse de nuevos colores en ninguna de sus intervenciones si puede utilizar lícitamente alguno de los ya empleados.

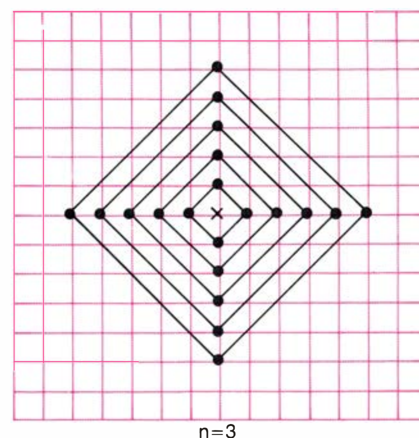
7. Whim (Capricho). En su libro *Gödel, Escher, Bach* (hoy editado en rústica por Vintage), con el que ganó el premio Pulitzer, Douglas R. Hofstadter introduce la noción de juego auto-modificante. Se trata de juegos en los que cada jugador, llegado su turno, tiene derecho a sacrificar su movimiento, a cambio de anunciar una nueva regla que modifica la estructura del juego. Estas reglas nuevas, electivas, se denominan meta-reglas. Las normas que modifican meta-reglas se llaman meta-meta-reglas, y así sucesivamente. Hofstadter da algunos ejemplos en ajedrez. En lugar de mover, el jugador de turno podría anunciar que, en lo sucesivo, no será lícito ocupar determinada casilla, o que los caballos deberán moverse de forma ligeramente distinta y, en general, decretar cualquier otra meta-regla tomada de una lista de alteraciones admitidas antes de empezar el juego.

La idea fundamental no es enteramente nueva. Ya en 1970, John Horton Conway propuso una original versión auto-modificante del conocido juego de Nim, versión que bautizó "Whim". El Nim es un juego bipersonal, donde los jugadores van retirando, por turno, una o varias fichas de un montón elegido por ellos, entre los diversos montones de fichas que ocupan el tablero. El número de fichas que contiene al principio cada montón es arbitrario; también lo es el número de montones. En la versión normal, el jugador que en su última intervención consiga retirar la última ficha del tablero gana la partida; en la versión *misère,* o versión negativa del juego, se actúa a la inversa, pues pierde quien se vea obligado a retirar la última ficha del tablero. Desde hace mucho tiempo, se conocen estrategias para realizar partidas perfectas; puede verse una en el capítulo dedicado al Nim, en mi libro *Scientific American Book of Mathematical Puzzles and Diversions* (Simon and Schuster, 1959).

La partida de Whim comienza sin que se haya establecido si la variante a jugar será la normal o la inversa. Empero, en cualquier momento del juego, cualquiera de los jugadores puede sacrificar su jugada y decidir si, en lo sucesivo, se adoptará la versión ordinaria o la versión negativa. Tal "jugada de capricho" solamente se puede realizar una vez; decretada la versión, ésta es inapelable, y habrá de continuarse con ella hasta el final. Se sabe ya que en Nim la estrategia es idéntica para ambas versiones del juego hasta casi el final, por lo que podría parecernos que la estrategia del Whim será cosa fácil de analizar. Pruebe a jugar unas cuan-







"Circunferencias" de radios 1, 2, 3, 4 y 5, en "taxigeometrías" generalizadas

tas partidas, y verá que la cuestión no es tan sencilla como parecía.

Supongamos que somos los primeros en jugar en una partida de Nim, con muchos montones de fichas y muchas fichas por montón. Supongamos también que esta versión corresponde a pérdida propia en el Nim ordinario. Deberíamos entonces ejercer el derecho a pedir la versión misère, pues la posición de las piezas queda intacta y, desde ella, somos ahora ganadores. Supongamos, en cambio, que la posición inicial corresponde a victoria propia, y que somos los primeros en actuar. Está claro que no nos atreveríamos a realizar ninguna jugada de la estrategia vencedora, pues ello le daría, al contrario, oportunidad de ejercer su jugada de "capricho", y dejarnos en posición perdedora. Por consiguiente, deberíamos realizar una jugada que en el Nim ordinario conduciría a la derrota. Por idéntica razón, el contrario deberá hacer también una jugada de pérdida. Está claro que tan pronto un jugador dejase de realizar jugadas "perdedoras", el otro ganaría optando por la versión contraria.

Conforme el juego progresa y se aproxima al fin, alcanzando el punto de bifurcación de las estrategias normal y negativa, para ganar resulta imprescindible pedir cambio. ¿Cómo determinar el momento preciso? ¿Cómo podríamos averiguar desde el principio de la partida qué jugador se alzará con la victoria, suponiendo que ambos bandos actúen lo mejor posible? La estrategia de Conway es fácil de recordar, pero como hizo notar en cierta ocasión, resulta difícil de adivinar, incluso por personas versadas en la teoría del Nim.

Kenneth W. Abbott, un consultor neoyorquino sobre problemas de cómputo, me ha enviado una bonita generalización de la taxigeometría discreta que comenté el mes de enero pasado. Como en la taxigeometría, los puntos del plano no-euclídeo son las intersecciones de las líneas horizontales y verticales de una hoja de papel cuadriculado. En la generalización de Abbott, la "distancia" entre dos puntos cualesquiera es un número real, definido por $\sqrt[n]{x^n + y^n}$, donde x se mide horizontalmente, y es el número de pasos que los separan en dirección vertical y n es un entero fijo dado.

Cuando n es 1 resulta la sencilla geometría taximétrica del pasado enero. En dicha geometría, todas las circunferencias son series de puntos equidistantes del centro de las mismas. Tienen la forma que vemos en la parte izquierda de la ilustración inferior de la página precedente, donde se han trazado las de radios 1, 2, 3, 4 y 5.

En el caso de ser $n = 2$, las circunferencias de los correspondientes radios han sido trazadas en la figura central. Observemos que las cuatro primeras circunferencias se reducen ahora a los cuatro puntos marcados sobre los ejes, estando el centro común a todas ellas en el origen de coordenadas. Tales circunferencias serán llamadas "triviales". Cuando n es 1, tan sólo la circunferencia de radio 1 es trivial; las restantes ya no lo son. Cuando n es 2, la quinta circunferencia es no-trivial. Existen ahora infinidad de circunferencias de cada tipo. En las triviales, el valor de pi es $2\sqrt{2}$; mas para las no-triviales, pi no tiene valor constante. En el caso de la quinta circunferencia, cuyo radio es 5, pi vale $(4\sqrt{10} + 2\sqrt{2})/5$.

Cuando n es 3, las cinco primeras circunferencias son todas triviales. En esta geometría, pi es $2^{(n+1)/n}$ para todas las circunferencias triviales.

Podemos proponer ahora una notable conjetura: todas las taxigeometrías generales con n mayor que 2 contienen, exclusivamente, circunferencias triviales. Como Abbott me ha hecho notar, esta conjetura es equivalente a la conocida por "último teorema de Fermat" y, por tanto, está pendiente de solución.

En la sección de febrero, mencioné la existencia de una sucesión de ocho números primos que se ajustaba a la pauta k, $k + 2$, $k + 6$, $k + 8$, $k + 12$, $k + 18$, $k + 20$, $k + 26$. Decía entonces que la única serie conocida de este tipo era 11, 13, 17, 19, 23, 29, 31, 37. John C. Hallyburton, Jr., que trabaja para la Digital Equipment Corporation, ha encontrado siete secuencias más, del mismo tipo. Los números iniciales de cada serie son:

15.760.091
25.658.441
93.625.991
182.403.491
226.449.521
661.972.301
910.935.911

A Mark Templer le debo haber caído en la cuenta de un error que cometí al mencionar cuatro primos correspondientes a un artículo suyo de 1980; el primero debió ser 31 y no 37, como se dijo. Otros lectores me hicieron observar otra equivocación. La sucesión de primos 7, 37, 337, 3337, ..., debió ser 31, 331, 3331,... Los números de esta serie son primos hasta el 333333331, que es producto de 17 y 19.607.843.

Taller y laboratorio

Los resaltos hidráulicos tienen un encanto especial, incluso los que se ven en el fregadero de la cocina

Jearl Walker

Cuando el chorro de agua de un grifo cae sobre un fregadero liso cuyo desagüe esté abierto, se forma un círculo en torno al punto de impacto. El radio de este círculo depende del volumen de agua transportado por el chorro, de modo que cuanto más débil es el caudal tanto menor es el círculo. Dentro del círculo, la profundidad del agua es menor que en el exterior; la transición entre ambas alturas se llama resalto hidráulico y se trata de una onda de choque estacionaria, que es la análoga hidráulica de las ondas de choque atmosféricas que crean los aviones supersónicos.

Aunque no caigamos en su cuenta, los resaltos hidráulicos aparecen en numerosas situaciones de lo más común. Así, pueden ocurrir en el agua que discurre por la calzada de entrada a un garaje o junto al bordillo de una acera; otros pueden contemplarse en canales de riego o pequeños arroyos. Pero los resaltos hidráulicos más espectaculares se dan en los estuarios de ciertos ríos cuando en ellos penetra la marea procedente del mar. Estos enormes resaltos hidráulicos, que se llaman barras de agua y también macareos, se desplazan aguas arriba a velocidades de hasta 12 nudos, tienen alturas de hasta unos seis metros y abarcan toda la anchura del río. La aparición brusca e inesperada de una barra de agua puede poner en grave peligro las embarcaciones que haya en el río.

De los tres tipos generales de ondas superficiales acuáticas, para la interpretación de los resaltos hidráulicos sólo importan las ondas de gravedad en aguas someras. En estas ondas, el movimiento está regido por la atracción gravitatoria que sufre el agua después de ser desplazada por primera vez. Su celeridad depende esencialmente de la profundidad del agua. Las ondas de gravedad en aguas profundas, que se desarrollan en la superficie de los océanos, no dependen de la profundidad y no desempeñan papel alguno en los resaltos hidráulicos, ya que éstos sólo se dan en aguas relativamente someras. Las ondulaciones del tipo que forman los insectos en las superficies de las charcas están gobernadas por la tensión superficial del agua, más que por la gravedad; estas ondulaciones, a veces llamadas ondas de capilaridad, tienen unas longitudes de onda relativamente reducidas (de algunos centímetros o menos) y no influyen en los resaltos hidráulicos.

Muchas veces hay que comparar la velocidad de una corriente de agua con la celeridad a la que se desplaza una onda de gravedad en aguas someras por la superficie de aguas tranquilas de la misma profundidad. Entonces, si el agua se mueve a mayor velocidad que las ondas, se dice que la corriente es supercrítica; en caso contrario, se dice que la corriente es subcrítica. Cuando ambas velocidades son iguales, se tiene una corriente crítica. Cuando una corriente pasa de supercrítica a subcrítica, se desarrolla un resalto hidráulico; esta transición es brusca y probablemente caótica porque durante ella la corriente se hace muy inestable.

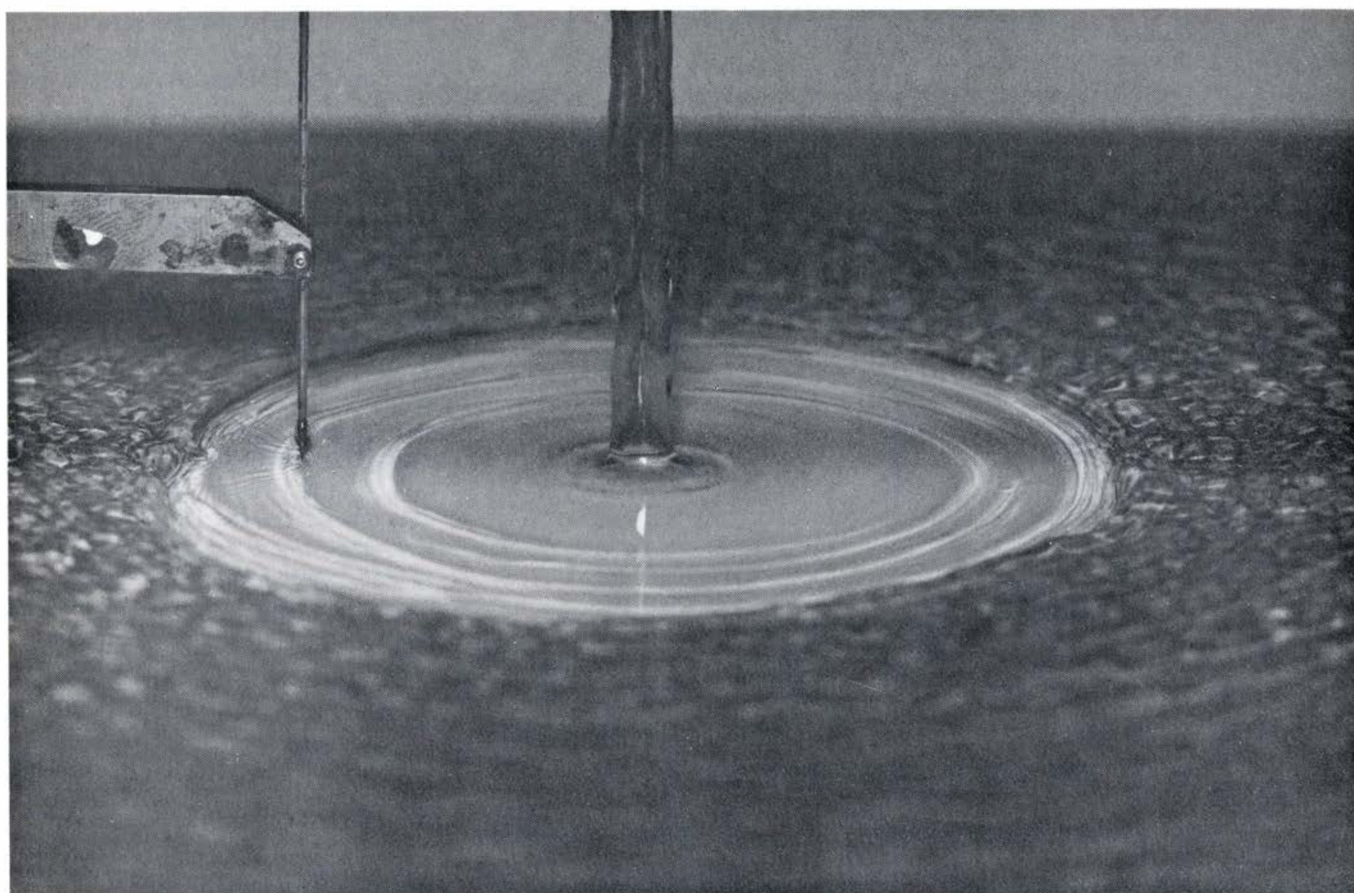
Cuando un obstáculo interfiere el movimiento normal de una corriente de agua, se generan ondas de tipo de gravedad en aguas someras. Si la corriente es subcrítica, las ondas pueden propagarse tanto aguas arriba como aguas abajo. En ciertos casos, puede haber una onda que permanezca estacionaria junto al obstáculo y aguas abajo del mismo. En efecto, si bien la celeridad de una onda depende fundamentalmente de la profundidad del agua, viene también condicionada por su longitud de onda; entonces, dentro de los límites de las longitudes de onda que un obstáculo puede engendrar, podrá darse una onda que se propague aguas arriba con una celeridad igual a la velocidad con que la corriente se desplaza aguas abajo. Además de mantenerse siempre en el mismo lugar, esta onda se verá reforzada continuamente por la interferencia del obstáculo con la corriente de agua. Las restantes ondas generadas por el obstáculo se propagarán alejándose del mismo, desvaneciéndose a la vez que se disipa su energía.

Cuando la corriente es supercrítica, no hay ninguna onda que se propague a mayor velocidad que la corriente y por ello ninguna viaja aguas arriba, siendo todas ellas arrastradas por la corriente. Si la corriente está pasando de supercrítica a subcrítica, puede formarse una onda estacionaria cuando la corriente atraviese el estado crítico. Entonces podrá darse la igualdad entre la velocidad de alguna onda que se propague aguas arriba y la velocidad de la corriente. A partir de una onda estacionaria como ésta puede desarrollarse un resalto hidráulico.

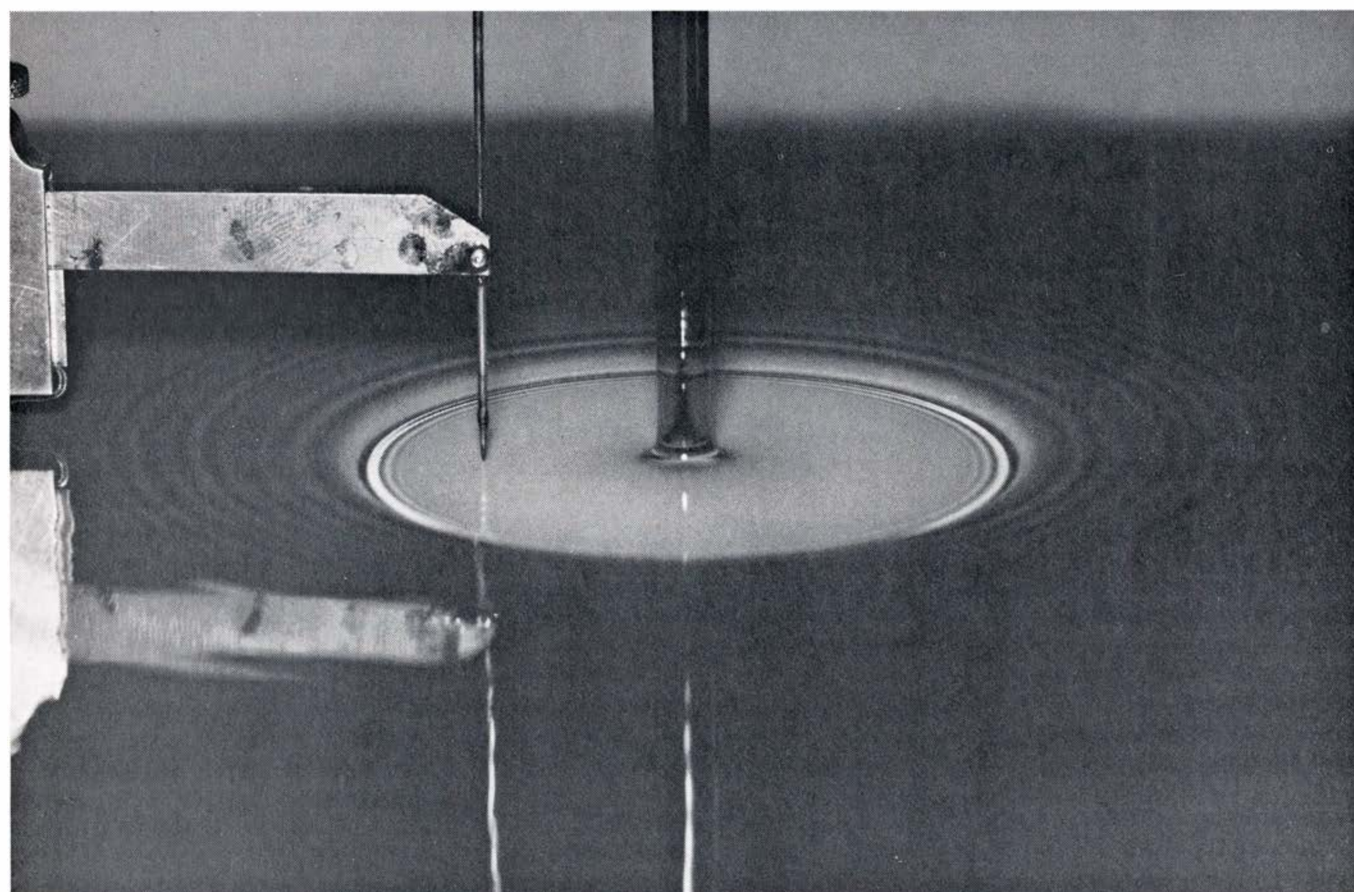
Un bajo existente en el lecho de un curso de agua nos proporciona un ejemplo de cómo un obstáculo puede crear una corriente supercrítica, una onda estacionaria y un resalto hidráulico en una corriente inicialmente subcrítica. Si el bajo es pequeño y produce sólo una resistencia reducida a la corriente, ésta sigue siendo subcrítica, pero más allá del bajo puede surgir una pequeña onda estacionaria. Si el bajo es mayor y produce más resistencia, la onda será mayor.

Un bajo aún mayor, y de superior resistencia, podría crear una onda con una zona relativamente poco profunda más allá del bajo y antes de la primera cresta. Como la celeridad de las ondas de gravedad en aguas someras depende de la profundidad del agua, el movimiento de las ondas será, en esta zona de poca profundidad, relativamente lento, obligando a que la corriente resulte en ella supercrítica. La primera cresta aguas abajo de la zona poco profunda devolvería la corriente a su estado subcrítico, a causa del aumento de profundidad. Si se mantiene la forma de la superficie del agua, la transición desde la zona poco profunda (corriente supercrítica) a la más profunda (corriente subcrítica) es un resalto hidráulico. Aunque un bajo no sea capaz de crear por sí sólo una onda estacionaria como ésta, sí podría desarrollarla a condición de que existiese después un segundo bajo que impidiese que una onda procedente del primero fuese arrastrada aguas abajo.

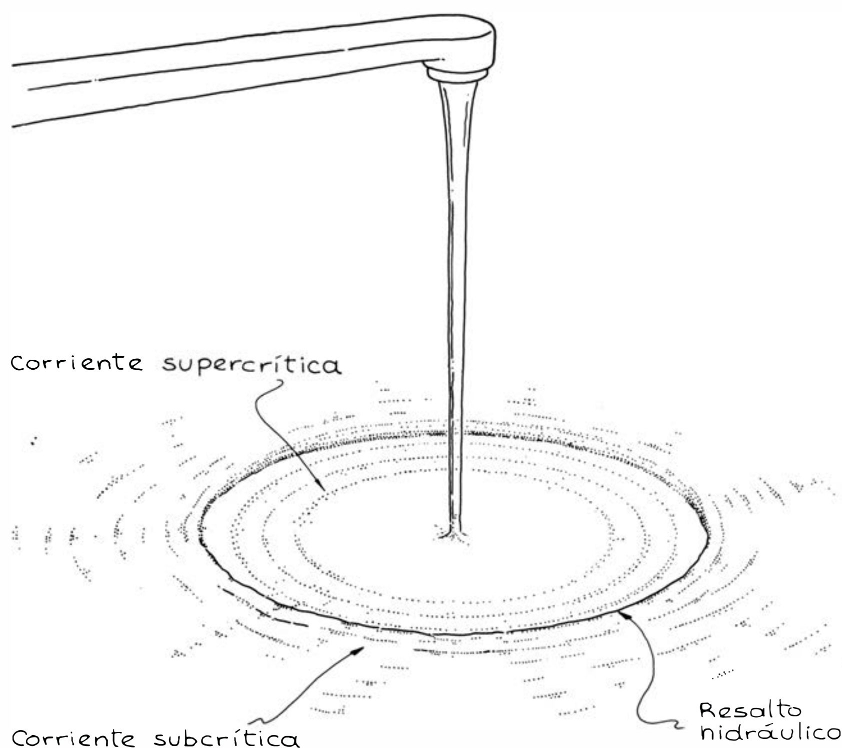
Normalmente la forma de los resaltos hidráulicos se clasifica de acuerdo con el llamado número de Froude, que da cuenta del estado de la corriente. El valor de este número es igual al cociente del cuadrado de la velocidad de la corriente entre el cuadrado de la velocidad de las ondas de gravedad en aguas someras propagándose en aguas tran-



Resalto hidráulico fotografiado por R. G. Olsson y E. T. Turkdogan



Efecto de añadir glicerina al agua



Detalles de un resalto hidráulico

quilas de la misma profundidad. Si la corriente es supercrítica, el número de Froude es superior a la unidad; si es crítica, el número de Froude es igual a la unidad y, si es subcrítica, inferior a la unidad.

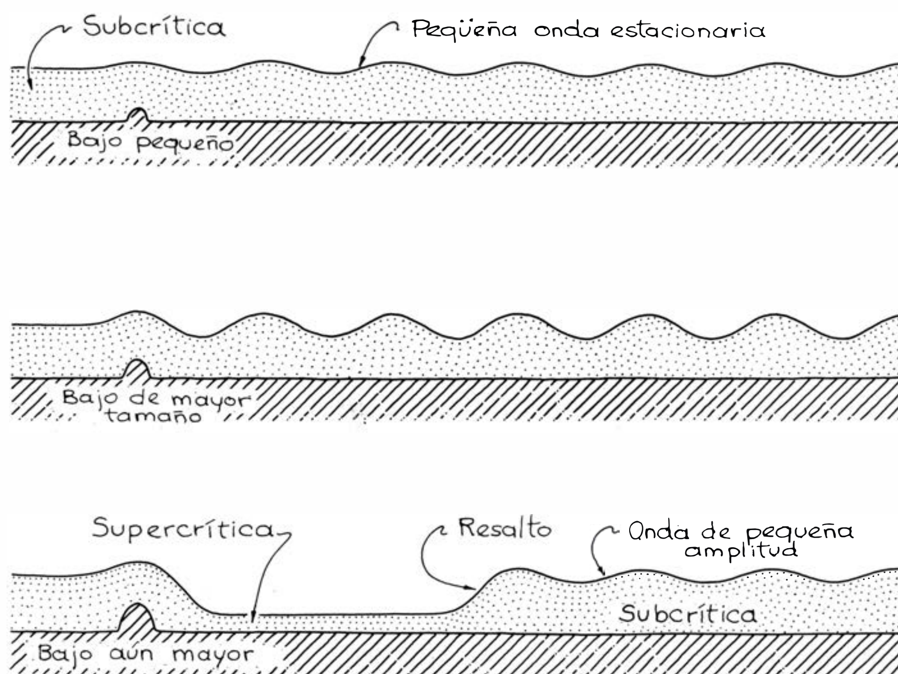
La forma de un resalto depende del número de Froude de la corriente supercrítica antes del resalto. (En principio la corriente sólo es inestable para un número de Froude de 1, pero la inestabilidad resulta tan caótica que se hace imposible ubicar exactamente dónde se da dicha condición. Esa es la razón por la que la clasificación se realiza empleando el número de Froude antes del resalto.) Si el número de Froude está comprendido entre 1 y 3, se dice que el resalto es ondular y se compone de una primera cresta grande y otras más pequeñas que la siguen aguas abajo; tras éstas, la superficie del agua aparece relativamente lisa. Para un número de Froude inicial comprendido entre 3 y 6, la transición entre los dos niveles del agua se hace más progresiva y no aparecen ondas posteriores; aquí se dice que el resalto es débil.

Si el número de Froude inicial vale entre 6 y 21, la transición origina ondas inestables de gran tamaño que pueden propagarse aguas abajo a distancias considerables del resalto; a causa de estas ondas irregulares, este tipo de resalto se llama oscilante. Cuando el número de Froude inicial se sitúa entre 21 y 80, se crea un resalto estacionario, que carece de ondas destructivas; en este caso, aproximadamente la mitad de la energía cinética del agua afluente se disipa en la turbulencia del resalto. Para números de Froude iniciales mayores, el resalto vuelve a hacerse irregular y agitado, disipa hasta el 85 por ciento de la energía cinética del agua y envía aguas abajo ondas potencialmente destructivas.

La velocidad de disipación de energía puede ser decisiva para el proyecto de una compuerta de presa o de otros sistemas de desagüe. A veces es necesario restar energía cinética a una corriente al objeto de evitar que las canalizaciones sufran daños. Para ello puede ser beneficioso colocar un obstáculo a través del fondo o una compuerta a través de la parte superior. El obstáculo debe diseñarse con gran cuidado, o bien debe ser ajustable de modo que pueda regularse el número de Froude y el resalto hidráulico no envíe aguas abajo ondas potencialmente destructivas.

Que una compuerta cree o no un resalto hidráulico en un canal dependerá de la inclinación de éste. Para un volumen de agua dado que discurra por segundo a lo largo de un canal, la pendiente determina la velocidad de la corriente y la profundidad del agua. Estos dos parámetros se relacionan de modo sencillo. Así, cuando el canal es escarpado, la corriente es rápida y la profundidad, escasa. Cuando la pendiente es más gradual, la corriente es más lenta y el agua, más profunda. Por definición, un canal escarpado es aquel que desarrolla corrientes de agua supercríticas. Los canales críticos generan corrientes críticas y los canales graduales, corrientes subcríticas.

Si sobre el agua que discurre por un canal gradual se hace descender una compuerta de modo tal que la corriente que emerja por debajo del canal sea supercrítica, se formará un resalto hidráulico para devolver la corriente al estado crítico que normalmente le corresponde en dicho canal. Si la corriente supercrítica procedente de la com-



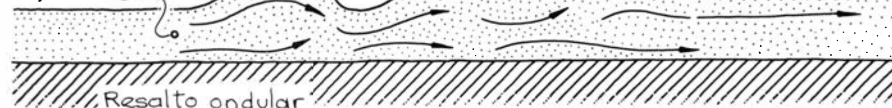
Corriente supercrítica producida por un bajo

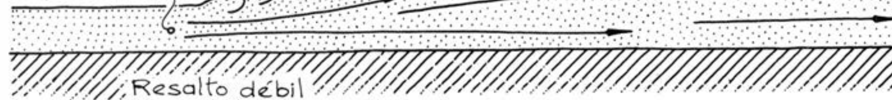
puerta se forma en un canal crítico, puede que no aparezca un resalto apreciable; aquí, la corriente pasa de supercrítica a crítica y permanece así, tornando inestable la superficie. Si la compuerta produce una corriente supercrítica en un canal escarpado, la corriente sigue siendo supercrítica y relativamente estable a lo largo de todo él.

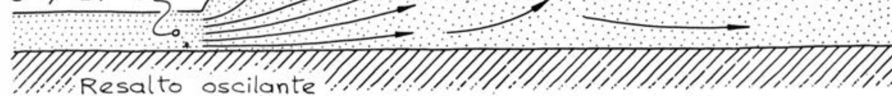
También puede surgir un resalto hidráulico cuando a un canal escarpado le sigue uno gradual, obligando a que la corriente de agua pase de supercrítica a subcrítica. En el canal escarpado el agua posee un número de Froude más bien elevado, y por ello permanece supercrítica y relativamente uniforme al ser bastante estable frente a las pequeñas perturbaciones producidas por obstáculos. Cuando el agua fluye sobre el canal gradual debe moverse más lentamente, de acuerdo con lo que es la corriente normal para tal pendiente, y el número de Froude disminuye aproximadamente hasta el valor para el que la corriente se hace inestable frente a las perturbaciones producidas por obstáculos. Además, se forman ondas y se desarrolla un resalto hidráulico para completar la transición a corriente subcrítica. Resultado de ello es una onda estacionaria que produce una variación repentina y espectacular de la profundidad de la corriente.

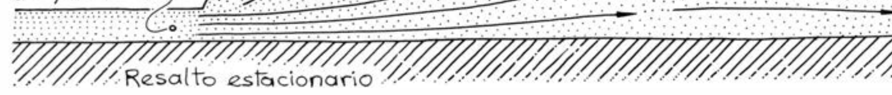
Este resalto puede que no se forme en la vecindad del lugar donde el canal escarpado se une al gradual. Su ubicación depende en parte de la pendiente del canal gradual. En efecto, cuanto mayor sea la pendiente del canal, tanto más paulatina será la disminución del número de Froude de la corriente y el valor crítico 1 se alcanzará en un punto más alejado aguas abajo; sólo entonces se formará el resalto. Si el canal gradual es sólo levemente inclinado, el número de Froude se reducirá antes y el resalto se formará en la proximidad de la unión de ambos canales, e incluso dentro del canal escarpado.

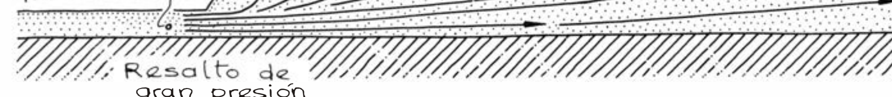
Cuando hace poco tiempo comencé a estudiar los resaltos hidráulicos, tuve que vérmelas con varias preguntas esenciales. ¿Por qué se presentan los resaltos hidráulicos? Como ya les he dicho, un resalto sirve como zona de transición para una corriente supercrítica que pasa a subcrítica; pero, ¿por qué es necesaria esa transición y por qué debe tener lugar bruscamente en un resalto? Si los resaltos aumentan bruscamente la profundidad de las corrientes, ¿por qué no se forman, por la misma razón, cuando la corriente es subcrítica? Y, por último, ¿por qué no hay resaltos en las transiciones desde una corriente supercrítica hasta otra también supercrítica, de número de Froude menor?









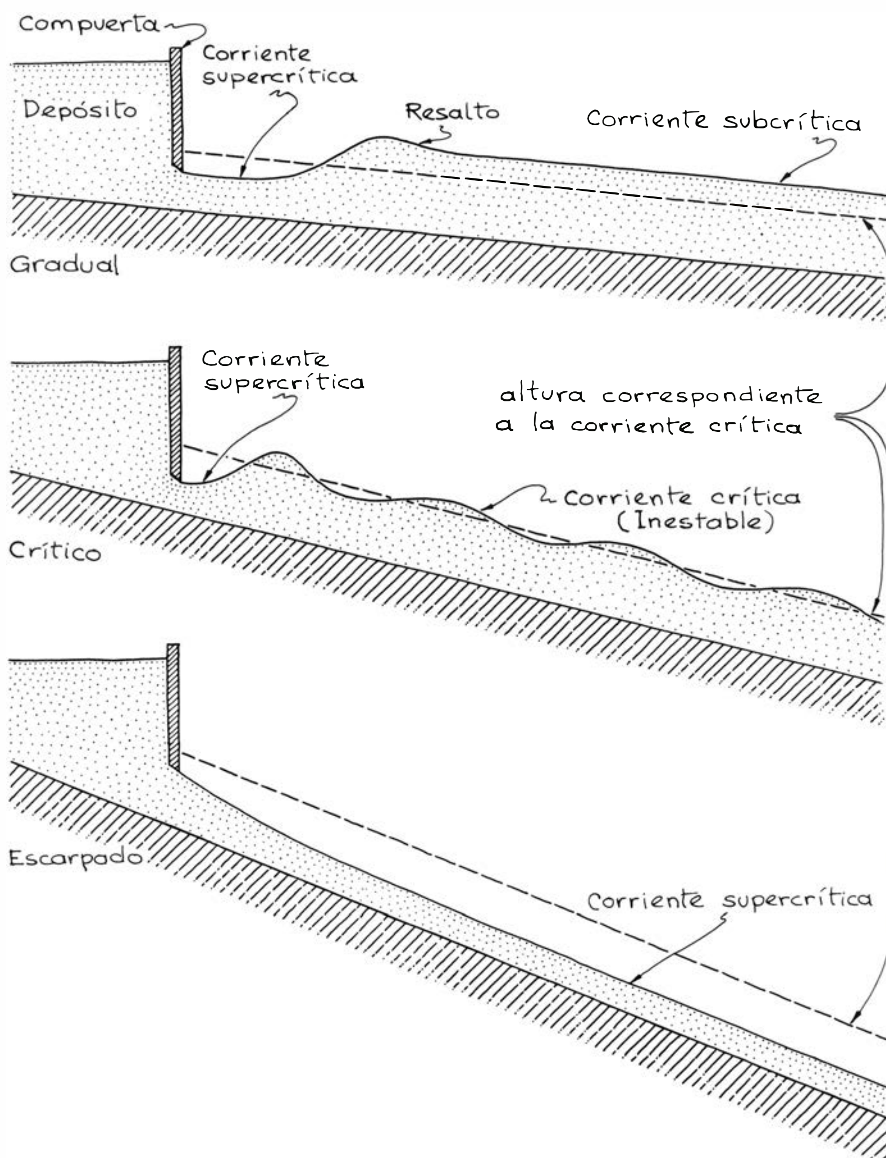


Tipos de resalto hidráulico

Permítaseme abordar estas cuestiones considerando una corriente supercrítica procedente de una compuerta o de un canal escarpado que se une a un canal gradual. La velocidad a la cual el agua puede fluir uniformemente por el canal gradual está regida por la pendiente y la rugosidad del mismo. El problema estriba en que el agua que penetra en el canal gradual se mueve con mayor rapidez que el agua que ya está dentro de él. Entonces, el agua que ya está en el canal producirá una resistencia que refrenará al agua entrante y, para que ésta pueda desplazarse aguas abajo, la profundidad deberá aumentar. Esta es la razón por la que una corriente de agua que fluye supercríticamente sobre un canal gradual empiece a moverse más lentamente y a hacerse más profunda.

Durante esta transición, que es más bien gradual, el número de Froude de la corriente disminuye hasta casi la unidad y la misma se hace sensible a las perturbaciones que pueda encontrar a lo largo del canal. Con sensible quiero indicar que basta con que la corriente encuentre una pequeña resistencia debida a un obstáculo para que aumente su altura considerablemente. Supongamos que la corriente (con un número de Froude próximo a la unidad) encuentra un pequeño bajo en el canal. La resistencia que este bajo opone a la corriente podrá ser muy pequeña, pero elevará la superficie hasta una altura comparativamente grande. Si la corriente encontrase el bajo teniendo un número de Froude diferente, la elevación consiguiente no sería tan espectacular.

Cuando la superficie del agua es impulsada repentinamente hacia arriba, se crean ondas. (Esto puede conseguir-



Efectos del aumento de pendiente en los canales

se también en un fregadero o en una bañera golpeando la superficie del agua, o elevando una mano sumergida por encima de la superficie.) Esta elevación completa rápidamente la transición a la profundidad mayor requerida por el canal gradual, y algunas de la ondas que así se crean quedan formando parte del resalto.

Hay varias razones por las que una transición de esa rapidez no tiene lugar entre dos corrientes subcríticas o entre dos supercríticas. Supongamos una corriente que fluye subcríticamente desde un canal gradual a otro canal más gradual todavía. En este segundo canal el agua tendrá que decelerar y hacerse más profunda para que pueda fluir a la velocidad dictada por las fuerzas que actúan sobre ella. En la transición entre ambas profundidades, sin embargo, no se formará resalto. Efectivamente, tan pronto el agua entrante comienza a moverse más lentamente y a hacerse más profunda, su número de Froude disminuye y se aparta del valor 1, valor crítico para el que la corriente es sensible. Y si esta corriente encuentra un obstáculo reducido durante la transición, la resistencia opuesta por el mismo a la corriente no producirá una elevación muy grande de la superficie del agua; el obstáculo perturbará la corriente, pero ésta recuperará en seguida la estabilidad.

Algo parecido ocurre cuando el agua fluye desde un canal supercrítico sobre otro, también supercrítico, pero de menor pendiente. Aquí disminuye el número de Froude, pero su valor permanece superior a la unidad y la corriente no es sensible a los pequeños obstáculos. La resistencia a la corriente podría originar un pequeñísimo aumento de altura, pero esta variación no sería espectacular y todas las ondas generadas se verían arrastradas aguas abajo por la corriente supercrítica.

Algunas de las propiedades de los resaltos hidráulicos pueden observarse sin más que colocar dos placas planas dentro del chorro de agua que sale de un grifo. Yo mismo hice la prueba de interceptar el chorro de agua que salía del grifo de mi cocina con una hoja de vidrio de ventana, manteniéndolo inclinado de modo que el agua se vertiera luego sobre un trozo de madera plano que sostenía apretado contra el extremo inferior del vidrio. El agua discurría por el vidrio sobre la madera y luego, desde el borde de la madera, hacia el fregadero. Las dos superficies inclinadas dispuestas de ese modo producen una corriente de agua diferente de la que fluye desde un canal escarpado a uno gradual, porque aquí el agua no está confinada entre paredes laterales; pese a ello, puede observarse la dependencia del resalto respecto del ángulo de inclinación.

Ajustando el ángulo formado por el vidrio y la madera me fue posible crear resaltos en casi todos los puntos de la madera. Para asegurarme de que el agua que se movía por encima del vidrio fluía supercríticamente, tuve que inclinar mucho el vidrio. Inclinando mucho la madera, aunque algo menos que el vidrio, el resalto aparecía más allá de la intersección del vidrio y la madera; con una inclinación muy pequeña (la madera incluso hubiera podido estar horizontal) se formaba el resalto en la intersección.

También experimenté con resaltos que se forman en canales inclinados, valiéndome de un sencillo montaje compuesto por una palangana de goma, dos perfiles acanalados de aluminio y un poco de cinta de plástico. Los canales metálicos estaban construidos de bandas de aluminio, que venían a medir un metro de longitud, más o menos, por varios centímetros de ancho. Doblé las dos bandas dándoles forma de perfil acanalado; atravesé una de ellas por la pared de la palangana, lugar del que previamente había recortado un poco de material. Los canales también hubiera podido hacerlos con trozos de canalón de tejado. Instalé luego una manguera de suerte que la palangana rebosara continuamente; así el caudal que fluía por el canal se mantenía constante.

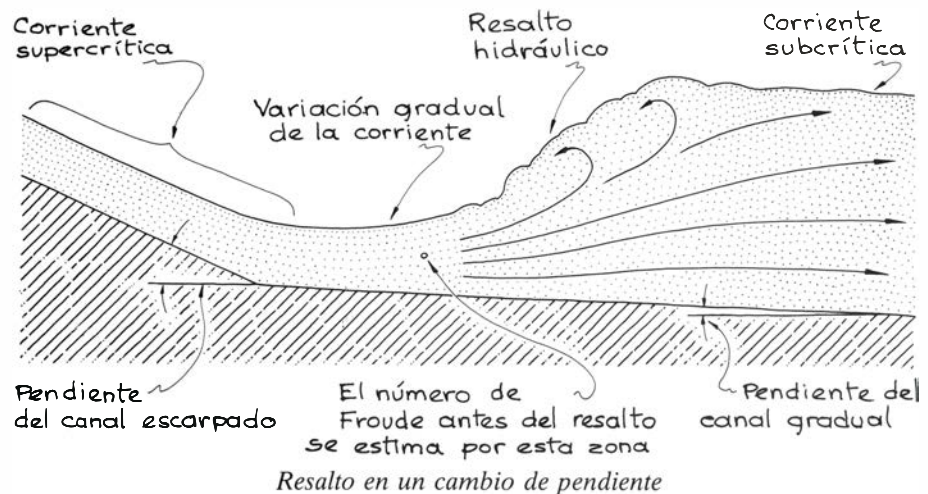
Fijé el segundo canal al extremo inferior del primero. Los ángulos de inclinación de los canales los ajustaba colocándoles debajo bloques de madera. El canal superior era escarpado y el infe-

rior, gradual. El agua discurría desde la palangana por el canal escarpado, luego por el canal gradual y, finalmente, por un desagüe del suelo del sótano. El canal lo fijé a la palangana con cinta adhesiva de plástico, arreglo que no resultó satisfactorio porque la cinta acababa aflojándose; sin embargo, me fue posible hacer casi todas las observaciones antes de que el montaje se viniera abajo.

Aunque crear resaltos me resultó sencillo, no me era posible decir de qué clase era cada uno. Constituye un montaje mejor construir los canales de plástico transparente para que el experimentador pueda contemplar los resaltos de costado; así sería más fácil determinar si la superficie del agua permanece lisa o si presenta ondas relativamente grandes.

Para estudiar los efectos de los bajos y de una compuerta en miniatura, retiré el canal inferior y coloqué obstáculos en el superior. Empleé de compuerta una placa metálica de casi la misma anchura que el canal. Ajustando la pendiente de éste y la profundidad de la placa, pude generar los distintos efectos que se producen en una compuerta. Como bajo puse un estrecho montículo de masilla a través de toda la anchura del canal. Ajustando la altura del bajo y la pendiente, pude controlar también la formación de resaltos.

Otra posibilidad para conseguir un bajo es construir una protuberancia lisa y gradual en el lecho del canal. El estudio teórico de lo que pasa cuando una corriente subcrítica encuentra una protuberancia como ésa sobre una pendiente gradual no es fácil. Puede ocurrir que la superficie del agua se eleve al pasar por encima de la protuberancia, si la corriente se hace supercrítica; y puede ocurrir también que la superficie se hunda al pasar el agua por encima de la protuberancia, si la corriente permanece subcrítica.

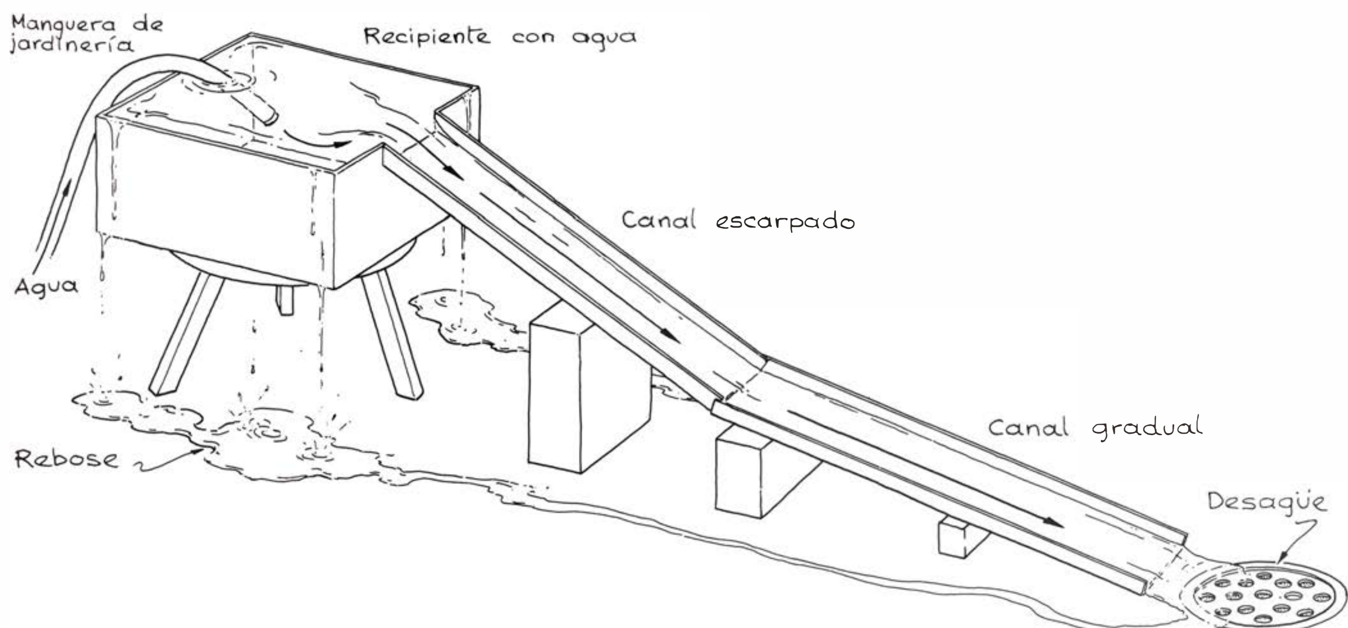


Resalto en un cambio de pendiente

El primero en dar cuenta del resalto hidráulico que se forma rodeando al chorro de agua que cae de un grifo fue Lord Rayleigh. En su trabajo titulado "De la teoría de grandes olas y barras de agua", publicado en 1914, Rayleigh organizó las ecuaciones que rigen el comportamiento de la energía cinética y de la cantidad de movimiento de una barra de agua. A manera de posdata incluyó sus sencillas observaciones del resalto hidráulico que se forma en un fregadero, haciendo notar que éste se rige por los mismo principios.

Hasta hace muy poco, y desde el trabajo de Rayleigh, los resaltos hidráulicos de los fregaderos de cocina recibieron poca atención. Uno de los estudios más interesantes acerca de los resaltos hidráulicos lo han publicado R. G. Olsson y E. T. Turkdogan, de la United States Steel Corporation. Lanzaron un chorro de agua sobre una placa plana circular mantenida perpendicularmente al chorro. El agua formaba un resalto sobre la placa.

El agua procedía de un depósito cuyo nivel se mantenía constante. Regulaban el diámetro del chorro mediante un orificio. Cuando sobre la placa se formaba un resalto, Olsson y Turkdogan medían la profundidad dentro y fuera de él introduciendo en el agua una escala de altura dotada de un nonio vertical y un índice de aguja. (Otro instrumento más ancho habría perturbado en exceso la corriente, destruyendo el resalto circular y aumentando la profun-



Montaje para estudiar las pendientes

didad del agua en las proximidades del aparato de medida.) En la parte interior del resalto la profundidad media oscilaba entre 0,1 y 0,9 milímetros. En la zona exterior, la profundidad variaba entre 1 y 3 milímetros.

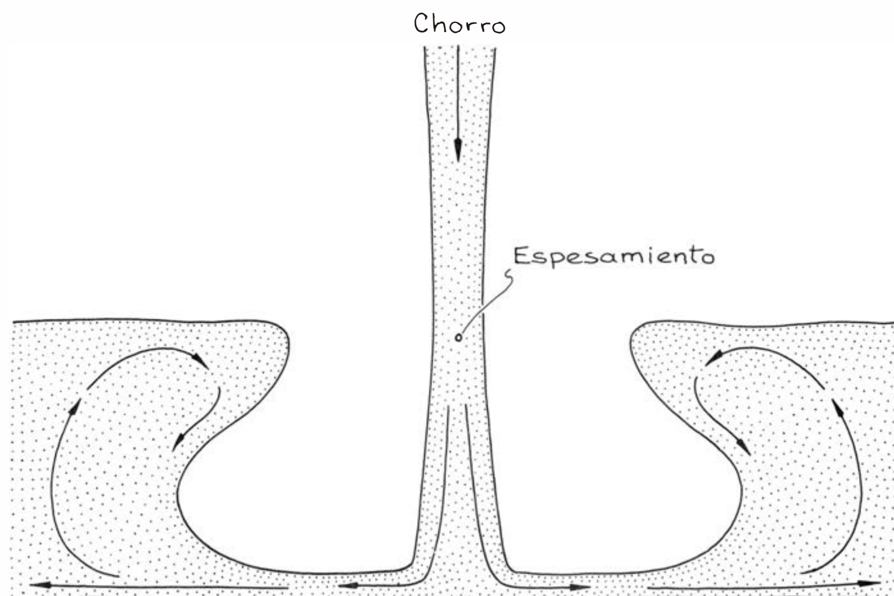
Filmando películas de gran velocidad (2000 fotogramas por segundo), Olsson y Turkdogan hicieron además una estimación de la velocidad de la corriente en la zona interior del resalto, utilizando para ello briznas de corcho que se movían en la superficie del agua. La velocidad resultó mantenerse aproximadamente constante hasta que el agua alcanzaba el resalto, punto donde comenzaba a disminuir. En otros experimentos sustituyeron el agua por fluidos más viscosos, resultando, en general, que, a mayor viscosidad, menor era el radio del resalto.

Yo mismo he generado resaltos hidráulicos con objetos diversos colocados en la trayectoria del chorro de un grifo. Un plato, una sartén e incluso el canto plano de un cuchillo de mesa crearon resaltos hidráulicos circulares completos o, al menos, partes de ellos. Hay que asegurarse el desagüe del fregadero o del recipiente, ya que si el agua se hace demasiado profunda, el resalto desaparece.

Cuando hice que el chorro cayera sobre una hoja de vidrio de ventana, introduje un poco de jabón en polvo en la zona de corriente supercrítica. Inmediatamente, el jabón fue arrastrado hacia el resalto, donde la turbulencia lo transformó en espuma. El jabón no disuelto se recogió precisamente al otro lado del resalto, donde lo dejaba el agua al disminuir la velocidad cuando la corriente pasaba a subcrítica.

Al crearse un resalto hidráulico en un fregadero de color oscuro o en una cazuela de hierro iluminados con luz fluorescente puede que se vean colores en la zona supercrítica. Yo mismo encontré bandas azules y amarillas (o anaranjadas) en torno al punto de impacto del chorro. Estos colores no se ven con luz incandescente o diurna. Las lámparas fluorescentes originan estos colores porque producen una luz que no es blanca, pese a lo que pueda parecer a la vista. La luz de una lámpara fluorescente tiene tres componentes fundamentales: la fosforescencia persistente de los luminóforos, o sustancias luminiscentes, que recubren el interior del tubo; la fosforescencia transitoria de otro tipo de luminóforos; y el espectro de emisión del mercurio excitado por la corriente de descarga a lo largo del tubo. La emisión uniforme de la fosforescencia persistente es más intensa en la región amarilla del espectro. La emisión de la fosforescencia transitoria se encuentra en la región del azul. Por eso, aunque el resultado medio parezca blanco, de hecho es un azul trémulo sobre un fondo constantemente amarillo.

Cuando un chorro de agua que cae choca con un objeto, se forma un movimiento ondulado camino del resalto hidráulico. Puesto que estas ondulaciones cambian la inclinación de la superficie del agua, modifican también la reflexión de la luz sobre el agua. A la vez que se trasladan desde la zona del impacto hacia el resalto, las ondulaciones reflejan unas veces luz amarilla y



Forma de resalto que aparece en una disolución de almidón de maíz y agua

otras veces una mezcla de amarillo y azul. Si el tren de ondulaciones es continuo, se verán círculos concéntricos de color amarillo y amarillo-azul alrededor del punto de impacto.

En una cocina pueden encontrarse muchos líquidos que sirven para generar resaltos hidráulicos si se vierten sobre un obstáculo plano y bien desaguado. Así, yo he interceptado chorros de aceite de maíz, vinagre, cerveza, almíbar, miel y una solución de almidón de maíz con una hoja de vidrio. Encontré que los líquidos de viscosidad relativamente baja exhiben unos resaltos hidráulicos similares a los que se forman con agua.

La miel, que estaba a la temperatura ambiente, resultó excesivamente viscosa para crear resalto hidráulico. Lo que pasaba es que, en vez de chocar con el vidrio y luego esparcirse hacia los lados supercríticamente, el chorro de miel se fundía lentamente con la delgada capa de miel que ya se encontraba en el vidrio. En el punto de impacto, un chorro fino se arrollaba sobre sí mismo en forma de rollo de cuerda; si el chorro era grueso, se movía de atrás adelante formando una cinta continua.

La solución de almidón de maíz pertenece a la clase de fluidos llamados no newtonianos, que se caracterizan por el hecho de que su viscosidad puede variar, no sólo por cambios de temperatura, sino también por cambios de tensión interna. Estos extraños fluidos los describí en el número de enero de 1979. La viscosidad de una solución de almidón se eleva inmediatamente cuando se la somete a tensión; al desaparecer ésta, la viscosidad torna de inmediato al valor inferior.

Preparé una solución de almidón en agua moderadamente espesa, de modo que fuera más espesa que el agua, pero no hasta el punto de que no pudiera fluir. Cuando vertí esta solución sobre una sartén, se formó un resalto hidráulico similar a los que se ven en agua. A la vez que subía el nivel del líquido en la sartén, el radio del resalto se contraía hasta un punto en que parecía desaparecer. A medida que disminuía el radio, el frente del resalto se hacía cada vez más empinado, pero en ningún momento manifestó turbulencia alguna; para el radio más pequeño de todos, el frente parecía ser cóncavo. En su caída, el chorro se espesaba al aproximarse al punto de impacto, chocaba con la solución que ya se encontraba en la sartén y luego se esparcía lateralmente para formar el resalto. Más hacia el exterior, a un radio de algunos centímetros, se evidenciaba un leve aumento de altura.

La extraña apariencia del líquido en la zona del impacto quizá se deba a la naturaleza no newtoniana de la solución de almidón. Lo que puede ocurrir es que, cuando la solución choca con la que ya está en la sartén, la viscosidad aumenta inmediatamente antes del punto de impacto e inmediatamente después del mismo. Resultado de ello es que el chorro colisiona con una superficie más bien rígida, aunque ésta sólo sea una solución de almidón; el fluido contenido en el chorro se proyectaba entonces horizontalmente en corriente supercrítica.

Tras el choque, el fluido sigue sometido a tensión, por lo que mantiene una viscosidad relativamente alta y, en consecuencia, el resalto presenta un radio reducido. El líquido fluye horizontalmente hacia el resalto y entonces un remolino transporta parte de la solución a la cima del resalto. El cerco muestra un saliente, posiblemente debido a la relajación de la tensión y consiguiente disminución de la viscosidad en esa zona. En cuanto el líquido es impulsado por el frente del resalto abajo y penetra en la zona de tensión creada por la corriente horizontal, la viscosidad vuelve a aumentar. El resultado es que el resalto presenta un frente cóncavo.

Cuando hice discurrir agua por la sartén, se formó un resalto hidráulico circular. Al inclinar la sartén, el resalto se deformaba, de modo que la parte de éste que quedaba más arriba se acercaba al chorro y la que quedaba más abajo se alejaba del chorro. El agua que se abría paso hasta la parte de arriba no se esparcía mucho, sino que, por el contrario, fluía hacia abajo en la cresta del resalto.

En los gases también pueden desarrollarse resaltos hidráulicos, y no sólo en los líquidos. Peter H. Hildebrand, del Servicio de Inspección de Aguas del estado de Illinois, fotografió uno de los casos más espectaculares de resalto hidráulico atmosférico y lo publicó en junio de 1977 en *Bulletin of the American Meteorological Society*. En aquella ocasión ocurría que una nube densa procedente de una tormenta de verano viajaba hacia el norte, pasando por Chicago en dirección al lago Michigan. Cuando la nube, que viajaba a gran velocidad a unos 200 o 300 metros de altura, se encontró con aire menos denso, reventó formando lo que sin duda fue un resalto hidráulico, dotado de un frente seguido de varias ondas. Este resalto se hizo visible al ser la nube densa notoriamente más oscura que el aire menos denso circundante.

Libros

Neurolingüística y memoria en Luria e historia reciente de la física nuclear

Antonio Tordera, Josep Maria Tous y Luis Alonso

FUNDAMENTOS DE NEUROLINGÜÍSTICA, por A. R. Luria. Editorial Toray-Masson; Barcelona, 1980. Las investigaciones llevadas a cabo por A. R. Luria son una muestra, rigurosa, de la constante preocupación a lo largo del siglo XX por los problemas del lenguaje. No sólo en lo concerniente a sus estructuras y funcionamiento, sino también, y de una manera global, como estudio de la capacidad cognoscitiva del hombre y su relación con el medio físico y el entorno social. La filosofía y la ciencia contemporánea se han caracterizado por una hipersensibilización respecto al problema del lenguaje que, por otra parte, ha sido considerado como definitorio del hombre y lugar central de su reflexión desde los inicios de la cultura.

La forma más reciente que ha adoptado la cuestión o el lugar teórico en el que hoy se produce es el cerebro, sus procesos y constitución. A partir de ese "topos", el viejo tema de la relación entre pensamiento y lenguaje se ha reactualizado, incluso polémicamente, en especial desde la llamada biología de la comunicación y en general en torno a la conducta y libertad del hombre.

En cierta manera, era teóricamente fatal la citada actualidad. No sólo porque la epistemología científica se mueve aún dentro de lo que Kuhn llamaría el paradigma lingüístico, sino por la estrecha relación existente entre los problemas planteados desde la psicología, la fisiología y la lingüística. Ya Saussure, por ejemplo, cuando reflexionaba acerca del lugar que le correspondía a la lingüística en el panorama general de las ciencias, a principios del siglo XX, y en orden a constituir aquélla como disciplina autónoma, señalaba que dicha ciencia debía ser considerada como una parte de la semiología y ésta como una rama de la psicología social.

La investigación de A. R. Luria es una ejemplificación específica de dicha conexión, en el sentido de que su enfoque desde una metodología, la patología clínica, ha abierto un camino propio en el conjunto de cuestiones antes descritas. De hecho, *Fundamentos de neurolingüística* es el resultado de la actividad científica de este autor ante tales cuestiones, actitud que desde la década de los años veinte se sitúa en el ámbito de la psicopatología experimental, para centrarse –después de sus estudios de psicología del desarrollo de la década de 1930 y el silencio propio en torno a la Segunda Guerra Mundial– en la investigación del cerebro y el área de las funciones mentales superiores.

Un estudio bibliométrico reciente realizado por J. M. Peiró, C. Matéu y H. Carpintero [véase *Revista de Historia de la Psicología*, n.º 1, 1980] de la obra de Luria muestra cómo progresivamente este autor ha ido delimitando la neuropsicología como la ciencia de la organización cerebral de los procesos mentales del hombre en orden a constituirla como una nueva área científica que, según se aprecia en la lectura del libro que nos ocupa, es pensada originalmente como una ciencia interdisciplinar, a fin de garantizar al máximo la objetividad de las conclusiones.

A partir de 1950 la teorización y práctica de Luria se interesa ya claramente por las funciones reguladoras del lenguaje, pero sitúa su método en el estudio de lo que él llama los procesos reales de la producción y recepción del mensaje verbal, esto es, el lenguaje en relación con el cerebro, en un intento de superar los idealismos de la filosofía y la psicología anteriores. A tal fin, la clínica y en concreto la descripción y análisis de las patologías y alteraciones de los diferentes procesos psicológicos y de las funciones corticales constituyen la metodología de A. R. Luria. Todo ello, como indica el responsable de la excelente traducción española, J. Peña Casanova, para establecer pautas a seguir en la rehabilitación de las funciones alteradas.

Finalmente, el marco de referencia introductorio queda completado si recordamos dos hechos. En primer lugar, que los textos de Luria son conocidos tempranamente fuera de la Unión Soviética. 1929 puede ser una fecha orientativa con bastantes garantías, ya que en ese año participa en el Noveno Congreso Internacional de Psicología celebrado en New Haven con una colaboración con L. S. Vigotsky, y publica también en ese año "The combined motor method and its application for the investigation of afferent processes", en la revista alemana *Psychologische Forschung*.

En segundo lugar, el silencio bibliográfico en lo que respecta a Occidente, debido a los avatares de la Segunda Guerra Mundial y posteriormente al aislamiento impuesto por Stalin. Silencio que se palía con amplitud, afortunadamente, con las numerosas traducciones de su obra a partir de los años sesenta. Así, de 103 trabajos originales de Luria, 73 son posteriores a 1960.

El conjunto de dichas publicaciones responde siempre a la intención metodológica de A. R. Luria de superar las limitaciones que tanto la lingüística como la psicología presentan en las cuestiones concernientes a la explicación de cómo se produce el lenguaje y a la relación del mismo con el pensamiento. En este sentido, el primer capítulo del libro que nos ocupa es un resumen del problema de la comunicación verbal y un repaso de las diversas soluciones propuestas desde finales del siglo XIX; en concreto, a partir de los investigadores afines a la escuela de Wurtzburgo de estudio del pensamiento, especialmente O. Külpe, Ach y Bühler. (A pesar de todo, la influencia de este último ha sido decisiva en la formación de la lingüística más válida, caso del Círculo Lingüístico de Praga o los escritos del lingüista R. Jakobson.)

La explicación de los procesos entre pensamiento y lenguaje debe asumir por una parte su complejidad real y por otra la dimensión social, histórica, de los mismos. Tal explicación, en opinión de A. R. Luria, debe apoyarse en los trabajos de su maestro L. S. Vigotsky, de quien puede verse a este respecto su clásico *El pensamiento y el lenguaje* (1934), que sin haber podido desarrollar con plenitud su teoría ofrece entre otros el concepto de "lenguaje interior" –o intermedio entre ambos polos de la cuestión central–, sobre el cual deben volver las investigaciones posteriores.

La cuestión se ha replanteado recientemente a propósito de la generación del lenguaje como una construcción del modelo "sentido ⇄ texto" que, ya en el campo específico de la lingüística contemporánea, trata de superar y completar la explicación chomskiana. Luria resume las etapas fundamentales como una serie de tipo representaciones semánticas-estructuras sintácticas superficiales-posterior desarrollo morfológico, fonológico y fonético.

En este proceso, el conjunto de "enlaces potenciales", que permite la representación semántica propuesta por I. A. Melchuk (sobre el cual pueden verse los magníficos estudios de Sebastián Serrano), es un esquema más satisfactorio que el del "lenguaje interno predicativo" de Vigotsky. A pesar de ello, concluye Luria, ni la lingüística ni la psicología esclarecen suficientemente los complejos procesos entre pensamiento y expresión verbal, por transcurrir éstos inconscientemente. Se necesitan nuevos métodos de investigación, que se concretan para Luria en el análisis neuropsicológico.

La ordenación del material clínico extraído de las historias de diversas alteraciones del lenguaje se ofrecen en este libro según dos grandes líneas antagónicas propuestas por la lingüística contemporánea: producción/recepción del mensaje verbal y construcción sintagmática/paradigmática de la expresión verbal; queda así estructurada la primera parte, mientras que la segunda es una amplia reconsideración sobre algunos tipos fundamentales de afasia.

La primera sección, por tanto, se dedica al análisis neuropsicológico de la formulación de la expresión verbal, cuya hipótesis central afirma que deben rechazarse las explicaciones estrechamente locacionistas y partir en cambio del supuesto de que la codificación de la expresión verbal se une a la adquisición y uso de los códigos del lenguaje e incluye una serie de factores psicofisiológicos. Cada uno de estos factores involucra distintos sistemas cerebrales interrelacionados entre sí. Por eso el déficit funcional de una zona concreta del cerebro inactiva a uno de estos factores particulares, dando como resultado la afectación de las formas de la actividad verbal que dependen de la integridad de este factor.

El estudio de las diversas patologías de la producción del lenguaje se agrupan en relación con el aparato sintagmático y, en un segundo apartado, con el aparato paradigmático. El abundante material expuesto le permite concluir a Luria que existen dos grandes bloques de patología del lenguaje: en las lesiones de "porciones anteriores del cerebro" se conservan, relativamente, la adquisición y uso de los códigos paradigmáticos; y queda afectada la capacidad de formular una expresión coherente y sintagmáticamente organizada, es decir, según el modelo lingüístico anteriormente expuesto, el paso del pensamiento al lenguaje coherente.

En cambio, en las lesiones de las "zonas específicas de la corteza cerebral posterior" (postcentrales, temporales y parieto-occipitales) se conserva la capacidad de producir una expresión verbal coherente y sintagmáticamente organizada y se altera la adquisición y uso de los códigos paradigmáticos.

El estudio de los procesos de codificación constituye un aspecto parcial que debe completarse con el análisis de la decodificación verbal, a la que se dedica la última sección de esta primera parte. Luria desborda allí los límites de los problemas lingüísticos para pasar a los del pensamiento verbal o la actividad conceptual considerada como un todo.

En este apartado se describen las alteraciones producidas en lesiones temporales (afasia sensorial); parietooccipitales (afasia semántica); afasias motoras (aferente y eferente); lesiones cerebrales y síndromes de alteraciones de la memoria, y lesiones frontales masivas. Así, Luria trata de demostrar que las lesiones cerebrales pueden afectar individualmente a cualquiera de los tres estadios del proceso inverso de la decodificación: identificación de los elementos lexicales, comprensión de las relaciones sintácticas que llevan, por último, al reconocimiento del sentido general del mensaje verbal.

La segunda parte del libro se dedica a una reconsideración de algunas formas clásicas de afasia compleja, es decir, las llamadas "afasia de conducción", "afasia motora transcortical" y, finalmente, la "afasia amnésica". Introduce cambios importantes; por ejemplo, a propósito de la afasia de conducción, que este autor considera como un trastorno que dependería en la mayoría de los casos de una imprecisión del análisis acústico-articulatorio, no tratándose en definitiva de una forma especial de afasia sino de un síntoma que aparece a expensas de la acción de diferentes mecanismos ligados estrechamente con los cambios de las tareas que se proponen al paciente. Asimismo, critica el punto de vista de la neuropsicología clásica sobre la afasia motora transcortical. El análisis de las alteraciones verbales de los pacientes demuestra que sólo repiten sin dificultad palabras aisladas y que la aparente integridad de su lenguaje repetitivo deja paso a un cuadro de profunda alteración en cuanto se examina su capacidad para repetir estructuras verbales más complejas.

Luria intenta superar, así, las habituales descripciones fenomenológicas de las afasias, buscando los factores parciales que dan lugar a las diversas formas básicas de las mismas. Reconoce, sin embargo, que la investigación no es definitiva, ya que pueden distinguirse tres fases teóricas en el estudio de las afasias: a) etapa en la que se creyó que se podía establecer una localización estricta en el cerebro de los procesos componentes del lenguaje; b) fase "neuropsicológica", que busca determinar los "factores primariamente alterados" como resultado de las lesiones focales a fin de comprender los mecanismos básicos que sustentan el lenguaje; c) tercera fase o "neurodinámica", de raíz pavloviana, que trata de comprender los síntomas básicos de la afasia en términos de sus "cambios parciales". Luria sitúa sus propias aportaciones en la segunda corriente, entendiendo que el análisis de los mecanismos "neurodinámicos" constituye un campo apenas delimitado.

A pesar de ello las bases objetivas para construir una nueva ciencia, la neurolingüística, quedan claramente expuestas en el presente libro. Si bien, defendiendo el carácter objetivo y necesario del método que propone, el mismo A. R. Luria no oculta una posible crítica que, sin embargo, no invalida el rigor y las sugerencias de su texto: la suposición de que el cerebro patológico no tiene por qué trabajar como el cerebro sano o normal. (A. T.)

NEUROPSICOLOGÍA DE LA MEMORIA, por Alexander R. Luria. Editorial Blume; Madrid, 1980. Representa, por su contenido, un homenaje póstumo en lengua castellana a la labor de este investigador ruso del que ya contábamos en nuestro idioma con otras excelentes obras. Homenaje, por cuanto esta obra representa la exposición sistemática y metodológica de un largo período de investigación del autor y su equipo de colaboradores, amén de explicitar el interés del mismo por los temas y planteamientos de la memoria en la psicología y la biología de hace veinte años para acá.

Además, consideramos que el libro tiene para los neurólogos y psicólogos hispanohablantes el valor de un programa de trabajo a llevar a término, lo cual, unido a la calidad de la información que explica, convierte a este libro en un clásico de la neuropsicología, por lo que el relativo retraso de su publicación en castellano no resta interés a su lectura. El lector podrá encontrar en la obra no sólo el excelente resultado del trabajo interdisciplinar, sino, también, las ventajas que se obtienen de la interacción entre la teoría y la práctica, cuando la contrastación empírica de la teoría se convierte en un medio útil de reeducación o terapia.

El libro está dividido en dos partes, cuya redacción corresponde, la primera, al año 1974 y, la segunda, al año

1976; pero la diferencia entre cada uno de estos apartados del libro no es sólo cronológica, sino fundamentalmente metodológica, ya que en la primera parte de la obra presenta la contrastación de las teorías sobre la memoria mediante el método experimental, utilizando como condiciones grupos de sujetos normales (control) y grupos de sujetos lesionados en su cerebro (experimental). En cambio, en la segunda parte del libro presenta, gracias a la utilización del método clínico basado en el estudio de casos individuales, la descripción de los síntomas y la estructuración lógica del síndrome, las relaciones entre la memoria y las restantes funciones psicológicas en base a la interrelación de las distintas áreas cerebrales.

Todavía hoy, después de haber transcurrido casi un lustro de la primera edición de esta obra de Luria, asistimos y participamos los psicólogos en la discusión de la definición conceptual de las tres actividades qûe configuran el tópico psicológico de la memoria. Todos aceptamos que la función mnemónica está definida de forma necesaria y suficiente por una actividad de impresión, por una actividad de conservación y por una actividad de reproducción de la información. El problema reside en que no todos los psicólogos estamos de acuerdo en el contenido conceptual que fundamenta estas denominaciones, ni, por tanto, en la definición operacional que de las mismas debe hacerse.

Así pues, la impresión consistiría, para los psicólogos que parten del modelo cognoscitivo, en una actividad registradora de la información que se realizaría a nivel periférico del sujeto y sólo implicaría los órganos sensoriales pertinentes. Esta posición permitió demostrar que la información impresionada era superior a la conservada posteriormente y manifiesta en la reproducción (paradigma del informe parcial); que la información impresionada tenía una duración temporal limitada, borrándose por sí sola a medida que se aumentaba el tiempo entre el final de la exposición del estímulo y el inicio del informe del mismo por parte del sujeto (paradigma de la amplitud de la demora); y que la información impresionada dependía de forma negativa del aumento de ruido que acompañase a la información estímulo (paradigma del enmascaramiento). Los primeros investigadores que aportaron datos empíricos que refrendaban esta posición fueron Sperling, Averbach y Coriell y Estes.

Por otra parte, los psicólogos que se basan en un modelo reduccionista (biológico, fisiológico o neurológico) consideran la impresión como una actividad que inicia la recordación y que se manifestaría en un cambio de la excitabilidad de las neuronas, provocando en las mismas una reactividad mayor al poner en marcha mayor número de sinapsis (hipótesis de los circuitos reverberantes). Esta posición permitió explicar la permanencia de la información estímulo en el organismo cuando éste ya había finalizado, y considerar de forma central y no periférica la actividad de impresión. Esta posición permitió, además, poner de manifiesto la importancia del sistema límbico en la impresión de la información (vigilancia-desinterés) y el papel del ácido ribonucleico en la impresión de la información (desencadenando la síntesis proteica o molecular neuronal que se considera la base del fenómeno de la habituación). Por consiguiente, investigadores como Laborit, Hyden y Luria están de acuerdo en considerar que la impresión de la información no tiene sólo el aspecto de una determinada duración en el interior del organismo, sino también la característica de consistir en una alteración o excitación específica de las células del sistema nervioso. Para Henri Laborit la interrelación entre los circuitos reverberantes y el desencadenamiento de la síntesis molecular neuronal aumentaría la permanencia de la información, considerándose ésta como un vestigio funcional de aquélla en el sistema nervioso. Para Luria, lo más importante consistiría, no en la duración de los vestigios funcionales de la información, sino en su carácter de excitación del sistema nervioso que le permite considerarlos desde la teoría de la inhibición-excitación.

La conservación de la información nos enfrenta con otro aspecto de la memoria. Todos los investigadores coinciden en la necesidad de distinguir entre una memoria a corto plazo y una memoria a largo plazo, y en que esta distinción es relativamente independiente de la impresión de la información e incluso de la reproducción de la misma. Al hablar de memoria a corto plazo o a largo plazo debemos entender que nos estamos refiriendo tan sólo a dos modalidades distintas de conservar la información. Para los psicólogos que se apoyan en el procesamiento de información, la conservación a corto plazo está vinculada con la actividad que el sujeto está realizando con la misma (paradigmas de la exploración y el repaso), de tal forma que cuando estas actividades terminan, aquella información, que gracias a las mismas no ha pasado a constituir la memoria a largo plazo, desaparece. En cambio, la conservación de la memoria a largo plazo vendría determinada por la incorporación de la información en un sistema organizado o estructurado que le conferiría una permanencia estable en el tiempo. Los diferentes tipos de simulación en la búsqueda y comparación mnemónica: serial exhaustiva, autolimitada y por direcciones constituyen un paradigma de la existencia de esta estructura dinámica en los sujetos; y, obviamente, el lenguaje o habla manifestaría que las estructuras fonéticas, sintácticas y semánticas constituyen una explicación de la mayor conservación de la información pertinente a las mismas. Por otra parte, para los psicobiólogos, los psicofisiólogos y neuropsicólogos la distinción entre memoria a corto plazo y a largo plazo tiene que ver con el carácter de excitación de los vestigios de la información en el sistema nervioso que denominamos trazo o huella de la memoria.

Para estos investigadores la conservación de la información no puede considerarse como un mantenimiento corto o largo, pero invariable de los trazos o huellas de la información, sino como una permanencia cambiable de los mismos relacionada con estados emocionales y cognitivos. Este carácter dinámico y cambiable de las huellas nuevas permite distinguirlas de las preexistentes en el momento de su impresión. Interesa aquí más el proceso de extinción, como proceso contrario al de conservación de la información, que la diferenciación entre distintas denominaciones de la misma conservación. Con todo, mantienen la distinción entre memoria corta y larga, por cuanto consideran que la primera es fundamentalmente una actividad de vigilancia o atención en la que está implicado el circuito de Papez, mientras que la segunda, o memoria a largo plazo, es una actividad de selección de la información aferente y la información pertinente a la evocación. Se considera, por tanto, la memoria a largo plazo como un sistema de codificación de la información que implica un propósito, una estrategia y una toma de decisiones que están vinculadas con las tareas fundamentales de los lóbulos frontales.

La reproducción de la información se entiende, en el modelo de procesamiento de información, como la evocación mediante la selección de respuestas de la información impresa y conservada en el sujeto; pero este proceso que debería entenderse como una actividad compleja de decodificación, se limita a la explicación de la transformación de la información aferente en unidades de instrucciones motrices fonético-articulatorias o musculares más o

menos compatibles con la ejecución habitual de los sujetos. Es evidente que para estos investigadores neo-asociacionistas el contenido de la información reproducida se explica tan sólo por el principio de la asociación, a pesar de haber admitido que, para la conservación de la información, debía considerarse la existencia de estructuras organizativas en el sujeto. Para los psicólogos reduccionistas, la reproducción de las huellas resulta de la superación de la inhibición por parte de aquéllas, lo que permite su fuerza (excitación) para ser reconocidas y evocadas.

En opinión de Luria, los diferentes estadios de procesamiento de información implicados en los procesos mnemónicos constituyen tan sólo la denominación de los distintos niveles en que se imprime, conserva y reproduce la información. Estos niveles son: sensorial, representacional y conceptual. Se diferencian entre sí por su mayor grado de complejidad progresiva, dado que cada uno manifiesta, respecto al anterior, un mayor número de conexiones, lo que aumenta la duración o permanencia de la información en ellos.

La memoria es un proceso activo de excitación, mientras que la extinción de la información u olvido consiste en un proceso activo de inhibición de la excitación por interferencia de excitaciones no pertinentes más fuertes. Esta interferencia, que puede ser tanto exógena (experimental o simplemente ambiental) como endógena (inercia patológica o fuerza negativa provocada por la lesión cerebral de los estereotipos preexistentes), es la mejor explicación de la pérdida de la información a nivel sensorial, representacional y conceptual.

A partir de la anterior posición teórica, Luria se plantea el olvido como aquel fenómeno que le permitirá establecer la relación entre la actividad funcional de ciertas zonas del cerebro y las diferentes funciones que delimitan la memoria. Así como para alcanzar un conocimiento de la memoria considera más adecuado investigar el olvido, entiende que para el estudio de la posible relación entre las bases neurológicas de la memoria y las funciones psicológicas de la misma conviene investigar sujetos con alteraciones clínicas de sus bases neurológicas y comparar los datos obtenidos con los registrados en sujetos normales. Realiza, además, un estudio cualitativo en el que compara las diferentes formas clínicas de alteración de la memoria que sirvieron para diagnosticar a los enfermos. Mediante la variedad de afecciones cerebrales (aneurismas, traumas y tumores profundos, de los lóbulos frontales, de la región temporal izquierda y de la región occipito-parietal izquierda), y por medio de la comparación cualitativa entre las mismas, determina que, a pesar de que la lesión produce en todos los casos un aumento de la inhibición, cada una de las zonas afectadas tiene un papel específico en relación con la memoria.

El resultado de la especificación de las zonas cerebrales vinculadas con la actividad mnésica consiste no sólo en un aumento del conocimiento de la topología cerebral, sino también en la distinción entre dos formas de influencia de las bases neurológicas en las funciones mnemónicas. A través del análisis de los enfermos estudiados demuestra que las lesiones que afectan a las zonas profundas del cerebro cumplen una función modal inespecífica que se manifiesta en alteraciones de la memoria corta o inmediata de carácter general; mientras que las lesiones correspondientes a las zonas gnósicas de la corteza sólo afectan a la memoria de una modalidad sensorial concreta (función modal específica) y no a la memoria para otras modalidades sensoriales.

Otro resultado de la determinación de las zonas cerebrales, relacionado con las funciones modales de las mismas, es el de la existencia de dos formas de alteración de la memoria. Según una, se altera tan sólo la reproducción de la memoria corta y, en cualquier caso, es posible observar que los sujetos mantienen la intención de recordar. Se manifiesta tanto en las afecciones de los sectores parieto-occipitales y temporales del hemisferio dominante y de forma modal específica, como en las afecciones de la zona límbica, talámica y subtalámica, aunque en este caso las alteraciones de la memoria tienen carácter modal inespecífico. Cuando las lesiones afectan masivamente a las zonas profundas de los hemisferios, la memoria queda afectada más gravemente que en los casos anteriores. En esas situaciones las alteraciones afectan sólo a la reproducción inmediata, pero el sujeto conserva su memoria a largo plazo. La otra alteración, que corresponde a la afección de los sectores frontales del cerebro, atenta contra la memoria a corto plazo y a largo plazo. Los sujetos no manifiestan intención de recordar ni son capaces de apreciar si su reproducción fue o no correcta.

Por último, gracias a la descripción de los síndromes que aparecen en las afecciones locales del cerebro, Luria presenta las relaciones que se establecen entre las alteraciones de la memoria, las alteraciones de la conciencia y las alteraciones de la actividad psíquica. Estas relaciones ponen de manifiesto que, a pesar de la interdependencia entre las distintas zonas del cerebro, cabe especificar la funcionalidad de cada una de ellas de forma relativamente independiente de las restantes, pero a la vez queda explícita la interrelación entre los diferentes tópicos psicológicos a nivel de los distintos sistemas funcionales del cerebro (J. M. T.)

FROM X-RAYS TO QUARKS, por Emilio Segré. Editorial W. H. Freeman and Company; San Francisco, 1980. Quizá sea Emilio Gino Segré el físico contemporáneo con más garra narrativa. Nadie contó con mayor viveza el resurgimiento de la física italiana que él en su biografía sobre quien fuera maestro y compañero suyo, Enrico Fermi. En la obra que reseñamos amplía su campo de consideración a todo el mundo occidental, aportando documentación de primera mano, cuando no su propia experiencia. Nacido en Tívoli en 1905, cambió su primer interés ingenieril por la física, doctorándose en 1928. Empezó su labor docente de adjunto de O. M. Corbino, en la Universidad de Roma. En 1930 recibió una beca de la Rockefeller Foundation para trabajar con Otto Stern en Hamburgo y con Pieter Zeeman en Amsterdam. De éste reconoce: "El efecto Zeeman se convertiría en una herramienta vigorosa para el estudio de la estructura atómica y, decisiva, para el principio de Pauli, el spin del electrón, el mecanismo de la emisión y otros muchos hallazgos". No menor es la admiración que guarda hacia Otto Stern, "uno de los físicos más eximios del período entre guerras. Hizo experimentos clásicos sobre la cuantización espacial, las ondas de Broglie y el momento magnético del protón. Experimentos todos ellos acometidos siguiendo el método del haz molecular".

En 1932 entra en el departamento de Fermi, donde colaboraría en las investigaciones sobre el neutrón. De 1936 a 1938 dirige el laboratorio de física de la Universidad de Palermo. Va entonces a Berkeley, centro a que se habrá de sentir siempre ligado, con el interregno del proyecto Manhattan. En 1959 compartió el premio Nobel de física con Owen Chamberlain por su descubrimiento del antiprotón. Ello nos retrotrae a uno de los momentos más creadores de la ciencia contemporánea, atrayentemente recogido en el libro. En 1928 Paul Dirac anunció que su combinación de las interpretaciones cuántica y relativista postulaba que la existencia real de una partícula (electrón) implicaba la existencia de su antipartícula (antielectrón, llamado también positrón), que estaría dotada de la misma masa aunque de

carga opuesta y momento magnético de sentido contrario. La predicción teórica de Dirac fue corroborada experimentalmente por C. D. Anderson, quien "no estaba familiarizado con la teoría de Dirac sobre el electrón y no sabía que predijera la existencia del positrón. Publicó una breve comunicación de su prueba experimental tan convincente que muy pronto persuadió a los físicos de que había observado un electrón positivo. Ello fue, ni que decir tiene, un claro triunfo de la teoría de Dirac".

El hallazgo de Anderson desencadenó un río de investigaciones en torno a las demás antipartículas, confiados los científicos en que la simetría de la naturaleza llegaba hasta esos extremos. La confirmación de la existencia del antiprotón, un sueño acariciado durante más de veinte años, resultó de tanta importancia para la física que la experimentación con nucleones (protones y neutrones) y la habilidad de Segré en los ensayos merecieron el Nobel. Así lo refiere aquí: "En 1955 el Bevatron de Berkeley conseguía una energía de 6 GeV, equivalente a unos 2 GeV en el centro de masa. Era lo mínimo requerido para producir un par protón-antiprotón, si es que existía realmente el antiprotón. O. Chamberlain, C. Wiegand, T. Ypsilantis y yo logramos demostrar su existencia de una forma convincente. Podría pensarse entonces ya en la existencia de la antimateria, e incluso de antimundos, por más que no sepamos hasta la fecha si esos antimundos existen".

Segré se mueve con plena seguridad en el intervalo histórico que va desde el nacimiento de la física atómica hasta 1960, más o menos (descubrimiento del electrón, rayos X, radiactividad –donde resume las vívidas experiencias que detalla en su obra sobre Enrico Fermi–, las teorías cuánticas y la relatividad especial; el núcleo, aceleradores y física de altas energías). A partir de entonces, los temas reciben un tratamiento más superficial. Aunque el hilo conductor mantiene ese difícil equilibrio entre el componente teórico del hallazgo y el diseño experimental que lo valida o refuta, sin perder tampoco el calor humano, la fuerza del trabajo en equipo y la solidaridad, la camaradería ya proverbial en todos los grupos. Por otro lado, la documentación gráfica aportada –facsímiles, aparatos, fotografías realmente históricas– enriquecen esta obra accesible al profano con cierto interés por la historia de su tiempo. Util también para el físico experto, quien podrá ir reconstruyendo la formación de las teorías, las aporías, los tropiezos y su superación final. (L. A.)

Bibliografía

Los lectores interesados en una mayor profundización de los temas expuestos pueden consultar los trabajos siguientes:

LIBERACIONES CATASTROFICAS DE RADIACTIVIDAD

REACTOR SAFETY STUDY: AN ASSESSMENT OF ACCIDENT RISKS IN U.S. COMMERCIAL NUCLEAR POWER PLANTS (WASH-1400). U.S. Nuclear Regulatory Commission, National Technical Information Service, 1975.

THE EFFECTS OF NUCLEAR WEAPONS. Dirigido por Samuel Glasstone y Philip J. Dolan. United States Department of Defense and United States Department of Energy, 1977.

TEORIA UNIFICADA DE LAS PARTICULAS ELEMENTALES Y LAS FUERZAS

UNITY OF ALL ELEMENTARY-PARTICLE FORCES. Howard Georgi y S. L. Glashow en *Physical Review Letters,* vol. 32, n.º 8, págs. 438-441; 25 de febrero de 1974.

WHY UNIFY? Howard Georgi en *Nature,* vol. 288, n.º 5792, págs. 649-651; 18 y 25 de diciembre de 1980.

RECONOCIMIENTO DEL HABLA POR MEDIO DE ORDENADORES

ON HUMAN COMMUNICATION: A REVIEW, A SURVEY AND A CRITICISM. Colin Cherry. The MIT Press, 1966.

TRENDS IN SPEECH RECOGNITION. Dirigido por W. A. Lea. Prentice-Hall, Inc., 1980.

ORIGEN DE LA INFORMACION GENETICA

THE GENERAL PRINCIPLES OF SELECTION AND EVOLUTION AT THE MOLECULAR LEVEL. Bernd Küppers en *Progress in Biophysics and Molecular Biology,* vol. 30, págs. 1-22; 1975.

THE HYPERCYCLE: A PRINCIPLE OF NATURAL SELF-ORGANIZATION, PART A, EMERGENCE OF THE HYPERCYCLE; PART B, THE ABSTRACT HYPERCYCLE; PART C, THE REALISTIC HYPERCYCLE. Manfred Eigen y Peter Schuster en *Die Naturwissenschaften,* vol. 64, n.º 11, págs. 541-565, noviembre, 1977; vol. 65, n.º 1, págs. 7-41, enero, 1978; vol. 65, n.º 7, págs. 341-369, julio, 1978.

PREBIOTIC EVOLUTION. Peter Schuster en *Biochemical Evolution.* Dirigido por H. Gutfreund. Cambridge University Press, 1981.

LAS ENVOLTURAS DE LAS NOVAS

STRUCTURE AND EVOLUTION OF CLOSE BINARY SYSTEMS. Dirigido por Peter Eggleton, Simon Mitton y John Whelan. D. Reidel Publishing Company, 1976.

THEORY AND OBSERVATIONS OF CLASSICAL NOVAE. J. S. Gallagher y S. Starrfield en *Annual Review of Astronomy and Astrophysics,* vol. 16, págs. 171-214; 1978.

THE SHELL AROUND NOVA DQ HERCULIS 1934. R. E. Williams, N. J. Woolf, E. K. Hege, R. L. Moore y D. A. Kopriva en *The Astrophysical Journal,* vol. 224, n.º 1, 1.ª parte, págs. 171-181; 15 de agosto de 1978.

INSECTOS FILTRADORES

THE ECOLOGY OF RUNNING WATERS. H. B. N. Hynes. University of Toronto Press, 1970.

THE ROLE OF FILTER FEEDERS IN FLOWING WATERS. J. Bruce Wallace, Jackson R. Webster y W. Robert Woodall en *Archiv für Hydrobiologie,* vol. 79, n.º 4, págs. 506-532, mayo, 1977.

AN INTRODUCTION TO THE AQUATIC INSECTS OF NORTH AMERICA. Dirigido por Richard W. Merritt y Kenneth W. Cummins. Kendall/Hunt Publishing Co., 1978.

FEEDING ECOLOGY OF STREAM INVERTEBRATES. K. W. Cummins y M. J. Klug en *Annual Review of Ecology and Systematics,* vol. 10, págs. 147-172; 1979.

FILTER-FEEDING ECOLOGY OF AQUATIC INSECTS. J. Bruce Wallace y Richard W. Merritt en *Annual Review of Entomology,* vol. 25, págs. 103-132; 1980.

NAVES DE GUERRA A REMO EN LA ANTIGÜEDAD

GREEK OARED SHIPS 900-322 A. C. J. S. Morrison y R. T. Williams. Cambridge University Press, 1968.

SHIPS AND SEAMANSHIP IN THE ANCIENT WORLD. Lionel Casson. Princeton University Press, 1971.

ANOTHER PUNIC WRECK IN SICILY: ITS RAM. Lucien Basch y Honor Frost en *International Journal of Nautical Archaeology and Undersea Exploration,* vol. 4, n.º 2, págs. 201-228; mayo, 1975.

THE HIGH SPEED CAPABILITIES OF ANCIENT BOATS. Sean McGrail y Ewan Corlett en *International Journal of Nautical Archaelogy and Undersea Exploration,* vol. 6, n.º 4, págs. 352-353; noviembre, 1977.

ENGINEERING IN THE ANCIENT WORLD. J. G. Landels. University of California Press, 1978.

PROTEOLISIS INTRACELULAR

MITOCHONDRIAL PROTEIN DEGRADATION. S. Grisolía, E. Knecht, J. Cervera y J. Hernández-Yago en *Processing and turnover of proteins and organelles in the cell,* dirigido por S. Rapoport y T. Schewe. Pergamon Press, 1979.

FATE OF PROTEINS SYNTHESIZED IN MITOCHONDRIA OF CULTURED MAMMALIAN CELLS REVEALED BY ELECTRON MICROSCOPE RADIOAUTOGRAPHY. E. Knecht, J. Hernández-Yago, A. Martínez-Ramón y S. Grisolía en *Experimental Cell Research,* vol. 125, págs. 191-199; 1980.

AUTOPHAGY OF FERRITIN INCORPORATED INTO THE CYTOSOL OF HELA CELLS BY LIPOSOMES. J. Hernández-Yago, E. Knecht, A. Martínez-Ramón y S. Grisolía en *Cell Tissue Research,* vol. 205, págs. 303-309; 1980.

PROTEIN DEGRADATION IN HEALTH AND DISEASE. Ciba Foundation Symposium, número 75. Excerpta Medica; 1980.

TALLER Y LABORATORIO

ON THE THEORY OF LONG WAVES AND BORES. John William Strutt, Baron Rayleigh, en *Scientific Papers;* vol. 6, 1911-1919, Cambridge University Press, 1920.

RADIAL SPREAD OF A LIQUID STREAM ON A HORIZONTAL PLATE. R. G. Olsson y E. T. Turkdogan en *Nature,* vol. 211, n.º 5051, págs. 813-816; 20 de agosto de 1966.

OPEN CHANNEL FLOW. Stephen Whitaker en *Introduction to Fluid Mechanics.* Prentice-Hall, Inc., 1968.

THE TIDES AS WE SEE THEM. Edward P. Clancy in *The Tides: Pulse of the Earth.* Doubleday & Company, Inc., 1968.